Analyse et prédiction de la relation séquence - structure locale et flexibilité au sein des protéines globulaires

Aurélie Bornot

UNIVERSITE PARIS.
DIDEROT (Paris 7)

INSERM UMR-S 665, Dynamique des
Structures et Interactions des
Macromolécules Biologiques (DSIMB)
INTS, 6 rue Alexandre Cabanel, 75739, Paris

U.F.R. Sciences du Vivant
Bâtiment Lamarck
35, rue Hélène Brion
75205 PARIS CEDEX 13 Case
7044

ECOLE DOCTORALE B3Mi

DOCTORAT

Discipline : Analyse de Génomes et Modélisation Moléculaire

BORNOT Aurélie

Analyse et prédiction de la relation séquence – structure locale et flexibilité au sein des protéines globulaires

Analysis and prediction of the sequence – local structure – flexibility relationships in globular proteins

Thèse dirigée par Alexandre G. de Brevern

Soutenue le 5 Novembre 2009

JURY

Pr. Fernando Rodrigues-Lima	Président
Dr. Frank Molina	Rapporteur
Dr. Annick Thomas	Rapporteur
Dr. Romano Kroemer	Examinateur
Dr. Sonia Longhi	Examinateur
Dr. Yves-Henri Sanejouand	Examinateur
Dr. Alexandre G. de Brevern	Directeur de thèse

A Jason,
A mes parents,

REMERCIEMENTS

Alexandre de Brevern est bien évidemment la première personne que je voudrais remercier pour m'avoir accompagnée dans cette thèse. Merci pour tes nombreux conseils tout au long de ces trois années. Merci également de m'avoir permis d'expérimenter l'enseignement ainsi que le travail en entreprise à travers le dispositif du Doctorat-conseil, malgré le temps que ces activités ont pris sur mon travail de thèse. Merci de m'avoir impliquée dans tant de projets parallèles. Et finalement, merci pour le temps passé à relire ce manuscrit.

Je remercie également Fernando Rodrigues-Lima d'avoir accepter de présider mon jury. Merci à Frank Molina et Annick Thomas-Brasseur d'avoir accepté d'être mes rapporteurs. Merci enfin à Romano Kroemer, Yves-Henri Sanejouand et Sonia Longhi pour le temps dédié à l'examination de ce travail de thèse.

Je remercie Catherine Etchebest pour sa présence et ses conseils tout au long de ces trois années de thèse mais également depuis mon master. Merci pour ton soutien dans la construction de mon projet professionnel, merci pour tes remarques avisées sur mon travail.

Guilhem Faure est également une figure importante dans cette thèse. Son acharnement au travail lors de son stage de master 1, a permis d'écrire deux publications. Merci Guilhem pour cette implication, pour ton enthousiasme et pour les quelques mois très agréables que tu as passé parmi nous. De même, merci à Olivia Doppelt-Azeroual, Manoj Tyagi et Amine Ghozlane.

Je remercie bien sur toute l'équipe DSIMB pour son soutien durant ma thèse et mais également bien avant lors de mon parcours universitaire. En particulier :

Je remercie Delphine Flatters pour toutes les explications qu'elles m'a fournit durant ma licence et mon master. Merci pour ta pédagogie. J'ai appris énormément de choses en biostatistique et en bioinformatique grâce à toi.
De même, je remercie Patrick Fuchs pour son dynamisme et ses supers cours en modélisation et en programmation. Merci aussi pour tes conseils durant ma thèse. Merci enfin pour le soutien technique que tu réalises au laboratoire depuis plusieurs années et depuis peu avec Pierre et Jean-Christophe. Un grand merci à tous les trois : sans vous, aucun calcul ne pourrait être lancé.

Merci à mes collègues de bureau. Joséphine, Lory et Marc, merci pour votre bonne humeur. Un grand merci à Joséphine en particulier. Merci pour avoir commencé chaque nouvelle journée de travail avec un grand sourire. Merci de m'avoir écouté. Et merci d'avoir partagé cette dure expérience de la rédaction d'un manuscrit de thèse avec moi.

Merci aussi à Paula d'avoir partagé de longues soirées de rédaction au laboratoire avec moi.
I also would like to thank Agnel. Thank you for your patience in listening to my stories in rough English. Thank you for your smile.

Un énorme merci à Jennifer. Jenny, merci pour ton soutien qui m'a été si précieux. Tu m'as tellement aidée dans tellement de domaines que je ne pourrais pas le décrire ici. Merci pour ta présence et ton amitié.

Par ailleurs, beaucoup d'autres personnes sont à remercier pour leur aide, leur soutien, leurs conseils tout au long de ses trois ans :

Je voudrais remercier Reine Note de tout cœur. J'ai été vraiment ravie de travailler avec toi lors de mon Doctorat-Conseil. Je ne te remercierai jamais assez pour cette expérience géniale, pour ton dynamisme, ta confiance, ton soutien, tous ce que tu m'as appris et tous tes conseils.

Merci à Thibault Colin, Corneliu Henegar et Laurent Bédouet, mes premiers directeurs de stage, qui m'ont fait confiance et m'ont permis de réaliser mes premières expériences professionnelles. J'ai beaucoup appris avec vous. Merci aussi pour votre soutien lors de ma recherche de postdoc et d'emploi.

Merci à Barbara Filler, mon mentor pour le nouveau chapitre de la thèse. Merci pour votre engagement et votre dynamisme. Merci pour vos précieux conseils.

Merci à la fondation L'Oréal pour m'avoir attribué la bourse L'Oréal – UNESCO – Les femmes dans la science. Cette bourse est actuellement d'un grand soutien pour le développement de ma carrière.

Pardon à enfin Hélène, Sophie, JF, Aude, Jérome, Vanessa et Seb. Nous ne nous sommes pas beaucoup vu pendant ces trois ans. Il va falloir que cela change. Merci à mes parents pour leur soutien permanent. Sans vous ce manuscrit n'existerait pas. Merci à mes sœurs pour leur patience. Et merci à Jason pour ton soutien indéfectible, ta présence, tes encouragements, tes conseils et ton amour.

Et puis :

J'aimerais remercier les gardiens des locaux de l'INTS pour leur bonne humeur, leur patience et leur présence même à des heures indues. Merci pour les cafés.

Merci enfin à W.L. DeLano, concepteur de Pymol, sans qui je n'aurais pas pu faire toutes ces belles figures et que je dois donc citer : (DeLano 2002).

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

AUC : *Area Under the ROC curve* en anglais, voir ROC.

BP : Bloc Protéique.

Cα : Carbone alpha, carbone central des acides aminés, auquel sont attachés quatre substituants : un groupe amine basique (-NH2), un groupe carboxyle acide (-COOH), un atome d'hydrogène et un groupe nommé chaîne latérale (-R).

DSSP : *Dictionary of Secondary Structure of Proteins* en anglais, méthode d'assignation des structures secondaires.

HPM : *Hybrid Protein Model* en anglais ou méthode de la Protéine Hybride en français.

IC : Indice de Confiance.

KLd : mesure de divergence asymétrique de Kullback-Leibler, utilisée pour mesurer la dissimilarité entre la distribution des acides aminés en chaque position des fragments de structure locale et la distribution des acides aminés dans la banque de données.

MS : Mots Structuraux.

Neq : Nombre équivalent.

PDB : *Protein Data Bank* en anglais, ressource mondiale de stockage des structures protéiques.

PSL : Prototype de Structure Locale.

PSSM : *Position-Specific Scoring Matrices* en anglais ou Matrice de Score Position-Spécifique en français.

RBF : *Radial Basis Function* en anglais ou noyau radial en français.

RL : Régression Logistique.

RL_PSSM : Stratégie de prédiction des structures locales couplant la régression logistique à l'utilisation de PSSMs.

RL_seq : Stratégie de prédiction des structures locales couplant la régression logistique à l'utilisation de la séquence seule.

RMN : Résonnance Magnétique Nucléaire.

RMSD : *Root Mean Square Deviation* en anglais, distance euclidienne calculée entre les coordonnées de deux fragments superposés de façon optimale. Ce critère permet de savoir si deux fragments possèdent une même géométrie.

RMSF : *Root Mean Square Fluctuations* en anglais, mesure l'amplitude des fluctuations d'un Cα donné tout au long de la trajectoire par rapport à une position moyenne de référence.

ROC : *Receiving Operating Characteristics* en anglais, les courbes ROC permettent d'évaluer le taux de vrais positifs en fonction du taux de faux positifs pour chaque classe.

SCOP : *Structural Classification Of Proteins* en anglais, classifications de domaines protéiques en fonction de leur repliement.

SOM : *Self-organized maps* ou SOMs en anglais ou cartes topologiques de Kohonen en français.

SVM : Support Vector Machine en anglais ou Machine à vecteurs supports en français.

SVM_PSSM : Stratégie de prédiction des structures locales couplant les SVMs à l'utilisation de PSSMs.

SVM_seq : Stratégie de prédiction des structures locales couplant les SVMs à l'utilisation de la séquence seule.

TPT : Taux de Prédiction Théorique.

UP : Unité Protéique

1. INTRODUCTION

Many problems of modern biology are concerned with the detailed relation between biological function and molecular structure. Some of the questions currently being asked will be completely answered only when one has an understanding of the structure of all the molecular components of biological system and a knowledge of how they interact. (Levinthal 1966)

Les protéines sont des constituants essentiels du vivant. Ces molécules géantes peuvent comporter des dizaines de milliers d'atomes et représentent la moitié du poids sec des cellules (Levinthal 1966) (cf. Figure 1).

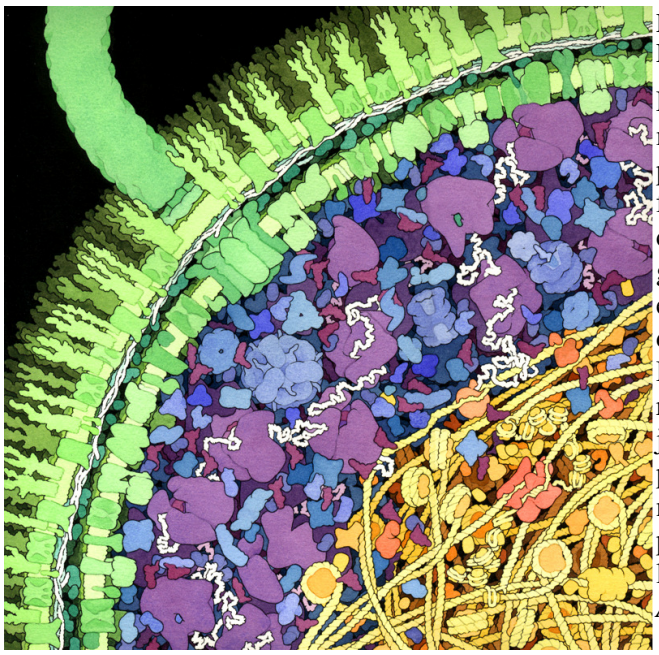

Figure 1. Illustration de la section d'une bactérie Escherichia coli par David Goodsell.

La paroi de la bactérie est constituée de deux membranes traversées par de nombreuses protéines membranaires et par un grand flagelle s'étendant vers l'extérieur (vert). Le cytoplasme est coloré en bleu et violet. Les grandes molécules violettes sont des ribosomes et les petites molécules marron en forme de L, des ARNt. Les brins blancs sont des ARNm. Des enzymes sont représentés en bleu. La région comportant le matériel génétique est en jaune et orange. Le long ADN circulaire est présenté en jaune et est enroulé autour de nucléosomes bactériens. Une ADN polymérase, impliquée dans la réplication de l'ADN, est représentée en rouge.
Auteur : David Goodsel

Ainsi, les protéines assurent des fonctions biologiques aussi importantes que variées : elles catalysent des réactions chimiques, assurent le transport d'ions et de petites molécules, participent à la régulation de l'expression des gènes, à l'empaquetage de l'ADN, au contrôle de l'apoptose ou encore à la réponse immunitaire. Un dysfonctionnement à leur niveau est susceptible d'entrainer ou de favoriser de graves pathologies comme la maladie d'Alzheimer, le diabète, des myopathies ou encore la mucoviscidose. Elles constituent donc une cible privilégiée pour le développement de nouveaux médicaments.

Ces macromolécules sont constituées d'acides aminés enchainés les uns aux autres dans un ordre précis porté par les gènes. Cet enchainement est nommé *séquence* ou *structure*

primaire. Depuis la fin des années 90, de nombreux programmes de séquençage de génomes ont été lancés et conduisent à une croissance exponentielle du nombre de séquences disponibles (Liolios et al. 2008). Toutefois, une part importante de ces séquences est difficilement exploitable en l'absence de protéines apparentées de fonction connues.

Or, la séquence en acides aminés se replie pour adopter une structure tridimensionnelle (3D). Cette structure, dite *tertiaire*, est déterminante pour la(les) fonction(s) biologique(s) des protéines. Sa connaissance est donc un atout majeur pour une compréhension de leurs propriétés fonctionnelles au niveau moléculaire. Dans le but de renforcer les connaissances de la communauté scientifique et d'aider à la compréhension de la relation séquence-structure-fonction, des consortiums de génomique structurale tentent d'organiser les efforts au niveau international pour déterminer la structure 3D d'un maximum de protéines différentes (PSI (*Protein Structure Initiative*) (Matthews 2007), RSGI (*Riken Structural Genomics/Proteomics Initiative*) et SPINE (*Structural Proteomics in Europe*) (Berry et al. 2006)). Cependant, en raison de difficultés techniques et du coût humain élevé nécessaire à la résolution expérimentale des structures, le fossé entre le nombre de séquences protéiques connues et le nombre de structures continue de se creuser.

Dans ce contexte, les méthodes de bioinformatique structurale ont un rôle particulièrement important à jouer et doivent fournir des méthodes alternatives permettant de réduire cet écart. Selon les travaux d'Afinsen, "toute l'information nécessaire pour obtenir la conformation native d'une protéine dans un environnement donné est contenue dans l'enchaînement des acides aminés" (Anfinsen 1973). Cette hypothèse est tout à fait fondamentale et fondatrice pour les méthodes de bioinformatique structurale visant à prédire la structure tertiaire d'une protéine à partir de sa séquence. Par ailleurs, selon Levinthal, la formation simultanée de petits noyaux structurés dans plusieurs régions d'une même chaîne polypeptidique initierait et accélèrerait le repliement (Levinthal 1968). Ainsi, cette seconde hypothèse, aujourd'hui considérée comme complémentaire de celle d'Afinsen, met en avant l'importance des interactions à courte distance et la formation de petites structures locales au cours du processus de repliement. Elle renforce donc la pertinence de la dernière génération des méthodes de prédiction des structures tertiaires : les méthodes dites *de novo*. Ces méthodes ne nécessitent pas l'existence de protéines "homologues" de structure connues. Elles reposent sur *(i)* l'identification de petits fragments de structure récurrents (ou structures locales), *(ii)* leur prédiction à partir de la séquence et finalement *(iii)* leur assemblage pour prédire une structure tridimensionnelle globale (Moult 2005). Le postulat est le suivant : même si tous les repliements possibles n'ont pas encore été observés, presque toutes les sous-structures ont

probablement été vues. Cette nouvelle vision mène à considérer les protéines comme un assemblage de fragments de structures récurrents dont certains pourraient être des sites d'initiation du repliement (Bystroff and Baker 1998; Fitzkee et al. 2005). Ces méthodes *de novo* remportent de grands succès depuis plusieurs années (Jauch et al. 2007).

Par ailleurs, pour aller vers une meilleure compréhension de la relation séquence-structure-fonction au niveau atomique, une autre caractéristique essentielle est à prendre en compte : les protéines ne sont pas des macromolécules rigides. Tout à l'inverse, elles sont flexibles et leur capacité à se déformer, étant au cœur des phénomènes d'interaction et de reconnaissance moléculaire, est souvent essentielle à la réalisation de leur(s) fonction(s) (Peng et al. 2007; Boehr and Wright 2008; Dunker et al. 2008). Ainsi, selon le modèle de *Sélection Conformationnelle*, bénéficiant aujourd'hui de plus en plus de crédit, une protéine non-liée existe en tant qu'ensemble de conformations dans un équilibre dynamique. La proximité d'un ligand peut favoriser la stabilisation d'une conformation peu peuplée et de plus haute-énergie, et ainsi modifier l'équilibre thermodynamique (James and Tawfik 2003; Boehr and Wright 2008; Lange et al. 2008). Dans ce contexte également, les méthodes de bioinformatique structurale peuvent être complémentaires des méthodes expérimentales. Par exemple, en l'absence d'information de structure, l'identification des régions les plus flexibles à partir de la séquence peut permettre de situer des régions potentiellement fonctionnellement importantes. Par ailleurs, connaître les régions flexibles d'une protéine peut être d'une aide précieuse lors de la résolution expérimentale d'une structure ou sa prédiction par des méthodes bioinformatiques.

Mon travail de thèse se situe principalement dans le cadre des méthodes de prédiction *de novo* des structures protéiques. Il s'articule autour de trois thèmes complémentaires : l'analyse de la structure tridimensionnelle des protéines, l'étude des structures locales et l'étude de la flexibilité des structures. Les objectifs de mon travail principal correspondaient plus spécifiquement aux deux derniers points et étaient *(i)* de créer une nouvelle méthode de prédiction des structures locales et *(ii)* de développer une méthode de prédiction de la flexibilité. Ces deux types de prédiction fourniront des informations très pertinentes pour l'optimisation des méthodes *de novo*.

Outre cette introduction et la conclusion, ce manuscrit s'organise en six sections.

Une première partie présente un état de l'art des méthodes de bioinformatique actuelles dédiées à l'étude des structure protéiques (*section 2 - Études Bioinformatiques Structurales des protéines*). Cette synthèse me permettra de replacer dans leur contexte différentes études

auxquelles j'ai pu participer au cours de ma thèse et portant sur des aspects variés comme l'assignation des structures secondaires, l'analyse et la prédiction des boucles, les implications structurales des mutations, l'analyse de surface protéique pour l'annotation fonctionnelle ou encore l'analyse des contacts au sein des protéines.

Les sections 3 et 4 présentent ensuite les méthodes d'analyse et de prédiction des structures locales (*section 3 - Caractérisation fine et prédiction des structures locales protéiques* et *section 4 - Nouvelle Stratégie de Prédiction des Structures Locales Protéiques*). La section 3 présente un état de l'art et développera plus particulièrement les travaux déjà menés au sein de mon laboratoire. La section 4 correspond à ma contribution à ce champ de recherche actif. Je me suis intéressée à la prédiction de structures locales relativement longues (11 acides aminés). L'analyse de longs fragments présente un intérêt majeur pour la prise en compte d'interactions à plus longue distance et pour aller vers la construction de modèles protéiques structuraux globaux. Toutefois, la prédiction devient également plus difficile du fait de l'augmentation de la variabilité de séquence. Ainsi, nous avons couplé des données évolutionnaires avec une méthode d'apprentissage sophistiquée pour proposer une méthode de prédiction des structures locales ambitieuse. Notre méthode propose un nombre limité de candidats structuraux pour une fenêtre de séquence et donne des résultats tout à fait compétitifs en comparaison des méthodes actuelles.

Les sections 5 et 6 sont dédiées à l'étude de la flexibilité des protéines (*section 5 – Flexibilité des structures protéiques* et *section 6 - Analyse et Prédiction de la Flexibilité Locale Protéique*). La section 5 présente un bilan du contexte scientifique actuel. La section 6 correspond à mon travail sur ce thème. Je présenterai notre questionnement sur la relation entre la qualité de la prédiction structurale et la flexibilité des structures. Cette réflexion a mené à l'élaboration d'une stratégie de prédiction de la flexibilité. Une comparaison de nos résultats tout à fait prometteurs avec ceux des méthodes existantes est présentée.

Les conclusions générales ainsi que les perspectives seront abordées dans la dernière section.

2. ÉTUDES BIOINFORMATIQUES STRUCTURALES DES PROTÉINES

Trois grandes classes de protéines sont généralement distinguées : les protéines globulaires, les protéines membranaires et les protéines fibreuses. Elles différent par de nombreux aspects. Les protéines globulaires sont solubles dans l'environnement aqueux du cytoplasme et le plus souvent compactes. Les protéines membranaires sont solubles dans la phase lipidique des membranes cellulaires. Enfin, les protéines fibreuses forment des fibres allongées et sont insolubles dans la cellule. Les différences fonctionnelles entre ces trois types de protéines sont également très importantes. Les protéines fibreuses ont un rôle mécanique et structural essentiel. Les protéines globulaires et membranaires sont porteuses de l'activité biologique de la cellule. Elles assurent par exemple des fonctions enzymatiques, hormonales, de récepteurs ou encore de transporteurs au sein organismes vivants. Par ailleurs, l'organisation structurale des protéines globulaires et membranaires est très différente et adaptée à leur environnement, aqueux ou lipidique. Dans ce travail, nous nous sommes spécifiquement intéressés aux protéines globulaires cytosoliques solubles dans l'eau. Elles sont caractérisées par un cœur hydrophobe et une surface hydrophile.

L'organisation structurale des protéines est hiérarchique. Quatre niveaux sont considérés :

- La *structure primaire* correspond à la séquence en acides aminés de la protéine dont elle détermine l'architecture et la fonction. Elle est le résultat de l'expression des gènes portés par l'ADN (Acide DésoxyriboNucléique). Le transfert de l'information génétique de l'ADN à la protéine n'est pas direct. L'ADN, dit codant, est d'abord transcrit en un ARN messager (Acide RiboNucléique). Cette ARN est ensuite traduit en une séquence protéique.
- La *structure secondaire* est une description des conformations locales régulièrement répétées de la chaîne polypeptidique. Les combinaisons courantes de structures secondaires sont les structures supersecondaires.
- La *structure tertiaire* correspond à la forme tridimensionnelle fonctionnelle de la protéine. Elle est déterminée par la séquence. Toutefois, des modifications post-traductionnelles (*e.g.,* glycosilations, phosphorylations ou encore clivages) peuvent également avoir un impact.
- La *structure quaternaire* est l'association de plusieurs chaînes polypeptidiques par des interactions non-covalentes ou des pontages covalents.

2.1 Structure primaire ou enchaînement des acides aminés

2.1.1 Acides aminés et Liaison peptidique

Les protéines sont des polymères linéaires constitués de 20 types majeurs d'acides aminés différents. Ces molécules élémentaires sont constituées d'un carbone central, nommé carbone α, auquel sont attachés quatre substituants : un groupe amine basique (-NH2), un groupe carboxyle acide (-COOH), un atome d'hydrogène et un groupe nommé chaîne latérale (-R) (voir Figure 2).

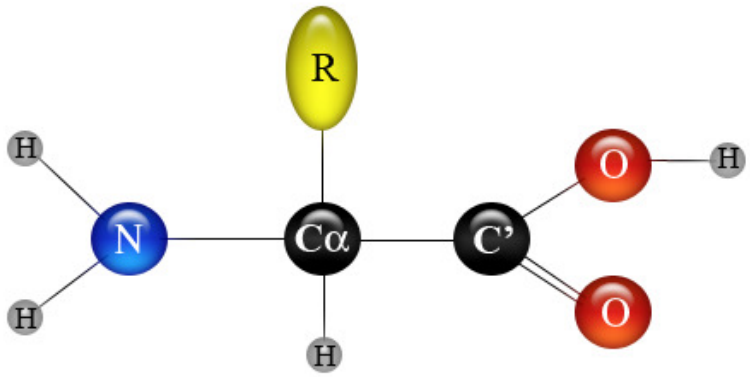

Figure 2. Schéma de la structure d'un acide aminé.
Le radical R symbolise la chaîne latérale caractérisant chaque acide aminé.

A pH (potentiel Hydrogène) dit physiologique de 7,4 et observé globalement dans les cellules biologiques, le groupe aminé est protoné et le groupe carboxyle déprotoné. Ainsi, à ce pH, la plupart des acides aminés sont des ions amphotères ou *zwitterions*.

Par ailleurs, à l'exception de la glycine, les quatre groupements portés par le carbone α sont différents. Les acides aminés sont donc chiraux et existent sous deux formes non superposables L (Laevo/gauche) et D (Dextro/droite). Toutefois, la forme L uniquement est impliquée dans la composition des protéines. En effet, seule la forme L est reconnue par les enzymes responsables de la polymérisation des protéines (Rawn 1990).

Chaque type d'acides aminés est associé à une chaîne latérale spécifique qui lui confère des propriétés physico-chimiques uniques. Le diagramme de Venn, présenté en Figure 3, propose une classification des acides aminés en fonction de ces propriétés (Taylor 1986). Ainsi, les acides aminés sont principalement caractérisés par leur hydrophobicité, leur hydrophilie et leur caractère acide ou basique. Ces propriétés sont essentielles pour l'architecture des protéines. En effet, les acides aminés chargés et hydrophiles sont majoritairement situés à la surface alors que les acides aminés hydrophobes sont généralement enfouis au cœur des

protéines. D'autres propriétés sont également très importantes telles la taille de la chaîne latérale, son caractère aromatique ou aliphatique. La glycine est le plus petit acide aminé car sa chaîne latérale est uniquement constituée d'un atome d'hydrogène. Cette propriété lui confère une fonction unique au sein des structures protéiques car elle peut s'intégrer à des espaces restreints excluant tout autre acide aminé. De même, la cystéine possède des propriétés remarquables : les groupes sulfhydryles de deux cystéines peuvent être oxydés en un pont disulfure et ainsi former une cystine. Ces ponts disulfures établissent des jonctions covalentes entre deux régions éloignées d'une même chaîne polypeptide ou des deux chaînes, stabilisant ainsi fortement des structures tertiaires ou quaternaires.

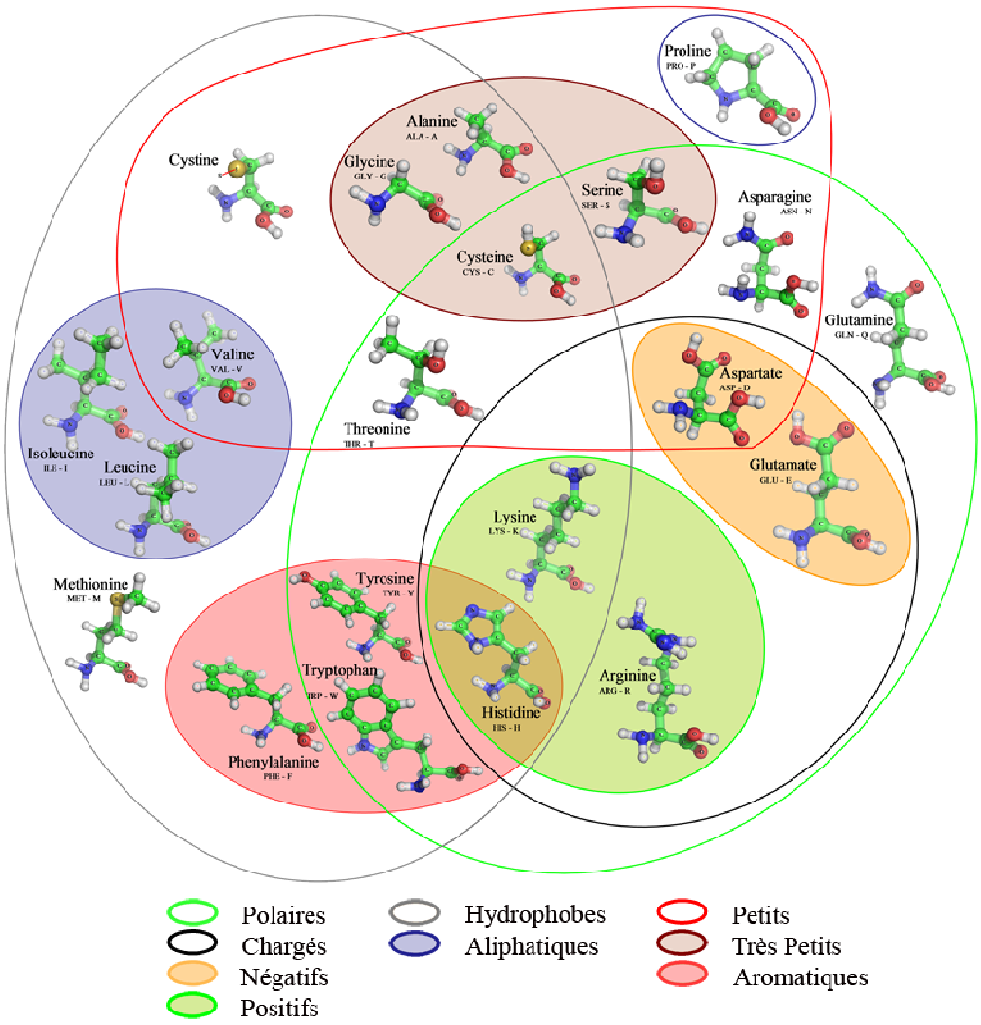

Figure 3. Diagramme de Venn regroupant les acides aminés en fonction de leurs propriétés physico-chimiques.

Cette représentation est une adaptation du diagramme proposé par Taylor (Taylor 1986). Figure adaptée de (Faure et al. 2008).

Au sein des protéines, les acides aminés sont enchaînés entre eux par une liaison peptidique (voir Figure 4). Cette dernière est créée grâce à la réaction entre le groupement carboxyle COOH d'un acide aminé et le groupement amine NH2 de l'acide aminé suivant aboutissant à la formation d'une liaison amide et d'une molécule d'eau. Les acides aminés ayant chacun perdu une molécule d'eau sont alors appelés résidus. La succession des atomes Cα-C'O-N constitue le squelette polypeptidique. Les groupes aminés et carboxyliques restés libres aux extrémités de la chaîne polypeptidiques sont respectivement nommés les extrémités N-terminale (chargée positivement) et C-terminale (chargée négativement).

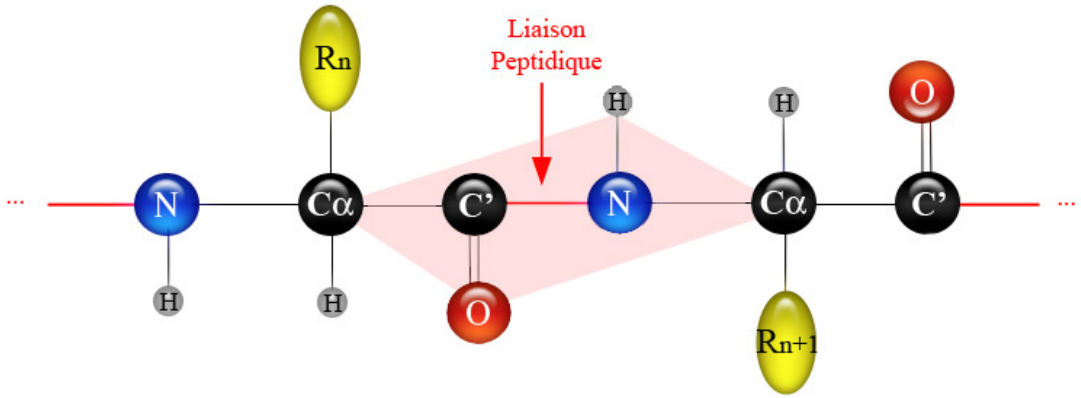

Figure 4. Liaison peptidique entre les acides aminés *n* et *n+1*.

Elle implique le groupe carboxyle de l'acide aminé *n* et le groupe aminé de l'acide aminé *n*+1. Les atomes $C\alpha_{Rn}$-C'O-NH-$C\alpha_{Rn+1}$ forment une unité peptidique.

Finalement, la structure primaire des protéines peut être décrite par la succession des résidus d'acides aminés en utilisant leurs abréviations en une ou trois lettres (voir les correspondances notées sur la Figure 3).

2.1.2 Détermination expérimentale : de la théorie du polypeptide au séquençage haut-débit.

Historiquement, la connaissance de l'organisation de la structure primaire est récente. La première détermination complète de la structure primaire d'une protéine, ou *séquençage*, a été réalisée pour l'insuline bovine par Sanger durant les années 1945-1955 (Sanger 1959). Ce travail capital lui vaudra le prix Nobel de chimie en 1958. Il établit définitivement la théorie selon laquelle les protéines sont des polypeptides constitués d'acides aminés enchaînés covalemment les uns autres dans un ordre précis. Avant cette preuve, plusieurs autres modèles

étaient discutés (Pauling and Niemann 1939; Hagen 2000) (cf. Figure 5). Il a notamment été suggéré que les protéines pouvaient n'avoir aucune structure moléculaire particulière et être uniquement constituées d'une suspension homogène de particules (modèle colloïdale). Wrinch également, avait développé le modèle des cyclols selon lequel les protéines étaient constituées d'un assemblage d'acides aminés sous forme de cyclopeptides (Wrinch 1940). L'un des modèles les plus répandues, plus proche du modèle du polypeptide, était celui de Bergmann et Niemann selon lequel les acides aminés étaient arrangés de façon périodique et apparaissaient de façon régulière le long de la chaîne protéique (Sanger 1959).

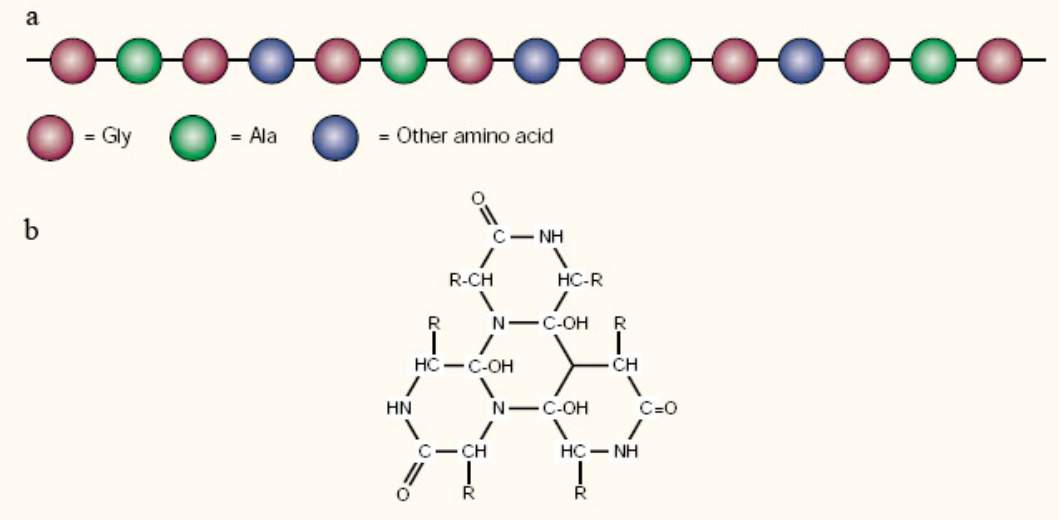

Figure 5. Théories alternatives pour la structure primaire des protéines.

a – Modèle de Bergmann et Niemann, enchaînement périodique des acides aminés. b – Modèle des clyclols soutenu par Wrinch. Figure adaptée de (Hagen 2000).

En imposant la théorie du polypeptide, Sanger a ouvert la voie de la biochimie des protéines moderne. Sa technique de séquençage s'appuyait sur la dégradation partielle des protéines en fragments suffisamment petits pour être séquencés directement par le jeu d'enzymes de restriction et par chromatographie. La première automatisation du séquençage fut développée par Edman et Begg (Edman and Begg 1967). La technique de la dégradation d'Edman permettait de séparer un à un les acides aminés de l'extrémité N-terminal de peptides. L'outil développé par ces biochimistes en 1967, nommé le *sequenator*, était alors capable de traiter un résidu par heure. De 1965 à 1978, les séquences protéiques résolues furent rassemblées par Dayhoff dans une série d'Atlas des séquences et des structures protéiques. Mais la croissance importante des données conduisit celle-ci à créer une version électronique accessible à la communauté scientifique. Cette première base de données de séquences protéiques fut nommée la PIR (Protein Information Ressource) (George et al. 1986). Durant les années 90, une nouvelle technique pour identifier très rapidement les séquences protéiques fit son

apparition : la spectrométrie de masse. Son principe est de fragmenter les polypeptides à analyser. Les protéines peuvent ensuite être identifiées en comparant la masse moléculaire de leurs fragments par rapport à des données dans des bases de données. Le développement de cette technique s'est effectué parallèlement à la croissance du nombre de séquences disponibles dans les bases de données et à l'augmentation de la puissance des ordinateurs (Henzel et al. 2003). La spectrométrie de masse est aujourd'hui largement utilisée dans les projets de protéomique à large échelle. Par ailleurs, avec l'augmentation de la résolution des instruments, le séquençage de nouvelles protéines est également devenue possible (Voet and Voet 1995).

Depuis la fin des années 90, les projets de séquençage de génomes à large échelle se sont multipliés. Aujourd'hui, ces projets sont de grands pourvoyeurs de séquences protéiques obtenues par traduction de séquences d'ADN codantes. Ces dernières sont soit générées expérimentalement, soit prédites par des méthodes bioinformatiques. Les génomes séquencés furent tout d'abord microbiens (*H. influenzae* et *M. genitalium* en 1995) puis très rapidement eucaryotes (*S. cerevisiae* en 1996). Depuis, plus de 1043 génomes ont été complètement séquencés et 3931 projets de séquençages sont en cours (http://www.genomesonline.org, Juillet 2009) (Liolios et al. 2008). Les projets de séquençage concernent les espèces bactériennes, les eucaryotes unicellulaires, les virus mais aussi des insectes, des plantes, des poissons des mammifères et bien sur l'Homme. Ils impliquent des consortiums internationaux d'instituts de recherche.

Actuellement, l'UniProt (Universal Protein Ressource) est la ressource mondiale centrale pour le stockage et l'annotation des séquences protéiques. Elle recense plus de 8,7 millions de séquences (http://www.ebi.ac.uk/uniprot/TrEMBLstats/). 99% de ces séquences sont des traductions ADN→Protéines provenant des projets de génomique.

2.2 Structure secondaire

Les protéines globulaires solubles dans l'eau sont caractérisées par un cœur hydrophobe et une surface hydrophile. Ainsi, au sein des protéines, la compaction de la chaîne polypeptidique est contrainte par ses propriétés géométriques et physico-chimiques. Deux contraintes principales sont observées :

- L'encombrement stérique des atomes des résidus limite les torsions possibles de la chaîne polypeptidique.

- Le squelette polypeptidique est polaire. Pour un repliement de la chaîne principale dans le cœur hydrophobe, ses groupements polaires doivent être neutralisés par la formation de liaisons hydrogènes. Cette neutralisation est rendue possible par la formation de structures secondaires.

Ainsi, les structures secondaires correspondent au deuxième niveau d'organisation de la chaîne polypeptidique. Elles décrivent des conformations locales adoptées par le squelette Cα-C'O-N.

2.2.1 Propriétés géométriques de la chaîne polypeptidique et liaisons hydrogènes

Suite à la polymérisation des acides aminés, la chaîne polypeptidique est constituée d'une succession d'unité peptidiques formées pas les atomes $C\alpha_{Rn}$-C'O-NH-$C\alpha_{Rn+1}$ (cf. Figure 4). Chaque unité peptidique est planaire et rigide. Cette propriété est la conséquence de délocalisations des électrons impliquées dans la liaison peptidique. La paire d'électron de la liaison C'=O du carbonyle se délocalise partiellement vers l'oxygène. De même, le doublet électronique libre porté par l'azote se délocalise partiellement vers le carbone du carbonyle. La liaison C'-N est donc double partiellement. La structure de la liaison peptidique est qualifiée d'*hybride de résonance*. Dans ce contexte, la conformation plane est la plus favorisée énergétiquement. Ainsi, l'angle de torsion, nommé oméga (ω), caractérisant la rotation possible autour de la liaison C'-N ne peut varier que de 1 % par rapport à des valeurs de référence de 180° (conformation trans), ou plus rarement de 0° (conformation cis) (Rawn 1990; Voet and Voet 1995).

De plus, l'oxygène étant plus électronégatif que l'azote, les électrons délocalisés de la liaison peptidique sont plus proches de l'oxygène. La liaison peptidique est donc polaire. L'oxygène du carbonyle est partiellement négatif et est accepteur d'hydrogène dans les liaisons hydrogènes. De même, l'azote amide est partiellement positif et peut jouer le rôle de donneur d'hydrogène. Cette propriété, fondamentale, est à l'origine de la stabilisation des structures secondaires par des liaisons hydrogènes établies entre les atomes du squelette polypeptidique.

Par ailleurs, des rotations des plans formés par les unités peptidiques sont possibles autour des liaisons N-Cα et Cα-C'. Les angles dièdres autour de ces deux liaisons sont nommés Phi (Φ) et Psi (Ψ) (voir Figure 6A). Toutes les conformations des angles Φ et Ψ ne sont pas autorisées à cause de l'encombrement stérique des atomes des chaînes latérales des résidus et

du squelette polypeptidique. Le biophysicien Ramachandran fut le premier à déterminer les régions autorisées en se basant sur des tripeptides (Ramachandran et al. 1963; Ramachandran and Sasisekharan 1968). Les résultats de ces recherches peuvent être résumés par le diagramme dit de Ramachandran représentant l'angle Φ en fonction de l'angle Ψ. Il met en évidence les conformations autorisées (cf. Figure 6B). Récemment, Ho et collaborateurs ont revisité ce diagramme et mis en évidence l'importance de clashes stériques non pris en compte initialement et de certaines interactions électrostatiques (Ho et al. 2003). Trois régions principales sont accessibles. Nous verrons que les conformations de la chaîne polypeptidique associées aux structures secondaires communément observées, sont comprises dans ces régions. Les angles de torsion observés pour la plupart des résidus au sein des protéines correspondent effectivement au diagramme de Ramachandran. La glycine et la proline sont les deux exceptions. La glycine, dont la chaîne latérale consiste uniquement en un atome d'hydrogène, peut adopter un plus grand nombre de déformations. Elle permet donc des changements de direction très serrés de la chaîne polypeptidique et a un rôle structural très important. A l'inverse, la structure cyclique de la proline impliquant les atomes du squelette polypeptidique, bloque l'angle Φ à -65°. Le nombre de conformations autorisées de la proline est donc plus limité.

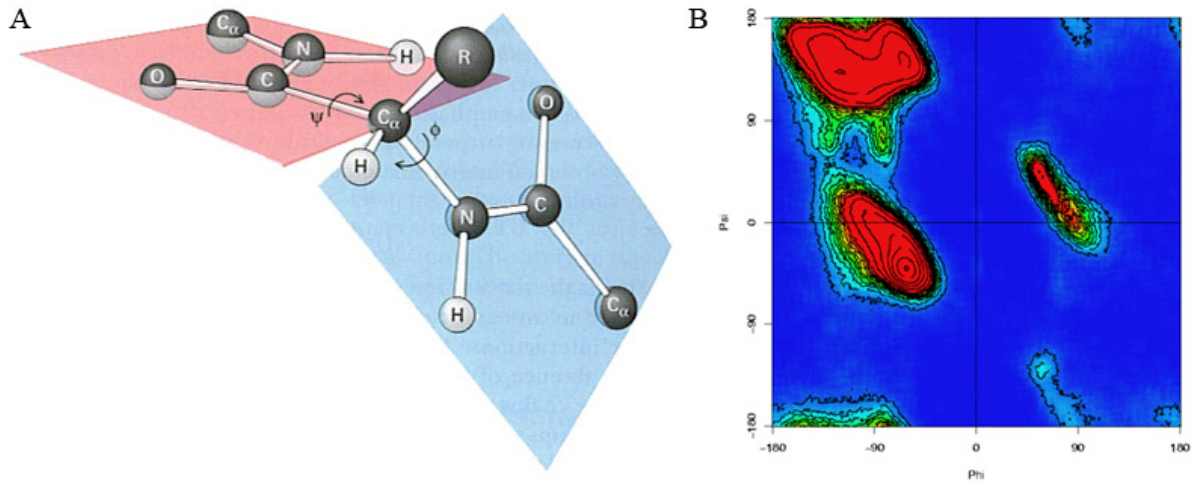

Figure 6. Angles dièdres Phi (Φ) et Psi (Ψ).

A – Définition des angles dièdres. L'angle Ψ correspond à la rotation du plan rouge autour de la liaison Cα-C'. L'angle F correspond à la rotation du plan bleu autour de la liaison Cα-N. Figure extraite du livre (Stryer 1996).
B – Diagramme de Ramachandran. Les couples d'angles (Φ,Ψ) (en degrés) autorisés sont représentés en rouge. Les deux plus importantes régions permises correspondent à des structures secondaires régulières de la chaîne polypeptidique. Les feuillets β sont en haut à gauche. Les hélices α droites sont en dessous. La troisième région rouge à droite correspond à des hélices gauches moins répandues. Figure extraite de (Benros et al. 2007).

2.2.2 Structures secondaires répétitives

Les structures secondaires répétitives définissent des conformations du squelette polypeptidique qui se répètent régulièrement. Les deux types de structures secondaires répétitives les plus représentées correspondent aux hélices α et aux feuillets β. Ces structures sont hautement stabilisées par des liaisons hydrogènes et sont énergétiquement très favorables à la chaîne polypeptidique. Elles représentent respectivement environ 1/3 et 1/5 des résidus. Elles furent prédites par Pauling, Corey et Branson en 1951 avant toute résolution expérimentale (Pauling and Corey 1951; Pauling et al. 1951; Eisenberg 2003; Offmann et al. 2007).

2.2.2.1 Les Hélices

Les hélices formées par le squelette polypeptidique sont caractérisées par des angles Φ, Ψ spécifiques, le nombre de résidus par tour de spire ou encore la distance entre deux enroulements consécutifs ou *pas* de l'hélice. La chaîne polypeptidique s'enroule, stabilisée par des liaisons hydrogènes, et les chaînes latérales se projettent vers l'extérieur.

L'hélice α droite est largement majoritaire. Elle représente 30% des résidus. Elle est caractérisée par des liaisons hydrogènes entre le groupe C'O du résidu *i* et le groupe NH du résidu *i*+4 (voir Figure 7). Ainsi, au sein de l'hélice, tous les groupes NH et C'O sont neutralisés par des liaisons hydrogène. Néanmoins, le premier groupe NH et le dernier groupe C'O aux extrémités restent libres. En conséquence, les extrémités d'hélices sont polaires et sont souvent localisés à la surface des protéines. Chaque tour d'hélice comporte 3,6 résidus. Les angles dièdres Φ, Ψ sont en moyennes égaux à -57° et -47° respectivement. Enfin, le pas de l'hélice est de 5,4 Å. Dans les protéines globulaires, la longueur des hélices peut varier considérablement et aller de 4/5 résidus à une quarantaine. La longueur moyenne est cependant égale à 14 résidus (±5) (Kumar and Bansal 1998).

D'autres hélices beaucoup moins fréquentes sont également observées. Les hélices 3_{10} et les hélices π représentent environ respectivement 4 et 0,02 % des résidus. L'hélice 3_{10} présente un enroulement plus serré que l'hélice α avec 3 résidus par tour et des liaisons hydrogène établies entre les résidus *i* et *i*+3. Elle a pour valeurs moyennes d'angles dièdres Φ= -49° et Ψ= -26°, son pas est de 6 Å. Ces hélices sont généralement courtes et constituées de trois (un tour) ou quatre résidus. Elles sont fréquemment observées aux extrémités des hélices α et

servent souvent de connexions entre deux hélices α. Les hélices π (4.4_{16}) correspondent à un enroulement plus lâche du squelette polypeptidique. Elles sont caractérisées par la présence de 4,4 résidus par tour et l'établissement de liaisons hydrogène entre les résidus *i* et *i*+5. Les pas de l'hélice π est de 5,2 Å. Cet enroulement crée un vide dans l'axe de l'hélice, trop petit pour laisser rentrer une molécule d'eau mais trop grand pour favoriser les interactions de van der Waals entre atomes de part et d'autre de l'axe. Les hélices π sont donc défavorables énergétiquement (Low and Baybutt 1952).

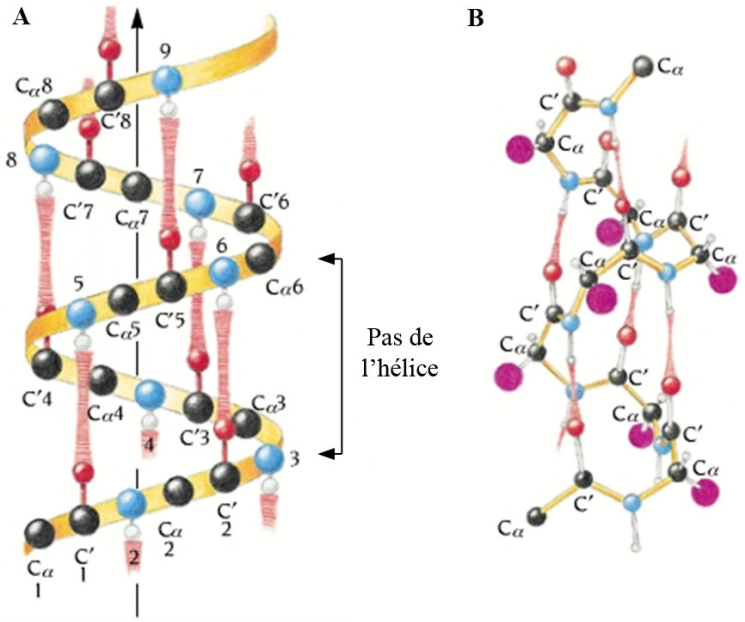

Figure 7. L'hélice alpha.

A – Les hélices sont souvent représentées de manière idéalisée sous la forme d'un ruban enroulé en spirale. Ici, sont également représentés les atomes du squelette polypeptidique et les liaisons hydrogène. Les atomes de carbone sont en noir, d'oxygène en rouge, d'azote en bleu et d'hydrogène en blanc. Les liaisons hydrogènes sont en rouge et striées.
B – Schéma du positionnement des atomes du squelette polypeptidique dans une hélice alpha. Le positionnement des chaînes latérales est indiqué en violet. Figure adaptée du livre (Branden and Tooze 1998).

Une relation dynamique pourrait exister entre ces différents types d'hélices. Les hélices 3_{10} et π ont été proposées comme intermédiaire de repliement des hélices α (Millhauser 1995). Par ailleurs, des transitions hélices α / hélices π ont été observées en dynamiques moléculaires (Lee et al. 2000; Armen et al. 2003).

Exceptionnellement, des hélices α ayant un sens de rotation gauche peuvent aussi être observées. Cependant, elles sont peu favorisées du fait de problèmes d'encombrements stériques.

2.2.2.2 Les Feuillets β

La deuxième structure secondaire majeure est le feuillet β. Ils représentent 20% des résidus. Contrairement aux hélices α, les feuillets β sont caractérisés par la formation de liaisons hydrogène entre régions distantes de la chaîne polypeptidique. Ils sont constitués de brins β alignés de façon à ce que des liaisons hydrogène puissent se former entre les groupes C'O de l'un des brins et les groupes NH du brin adjacent et inversement. Ces brins ont une structure étendue et sont en général longs de 5 à 10 résidus (Branden and Tooze 1998).

Deux conformations de feuillet β sont observées (cf. Figure 8):

- Les feuillets parallèles dans lesquels les brins β sont orientés dans des sens identiques. Les angles dièdres (Φ, Ψ) du squelette polypeptidique sont alors égaux à -119° et +113° en moyenne.
- Les feuillets antiparallèles dans les lesquels les brins β sont orientés de manière inversée. Les angles dièdres (Φ, Ψ) du squelette polypeptidique sont en moyenne -139° et +135°.

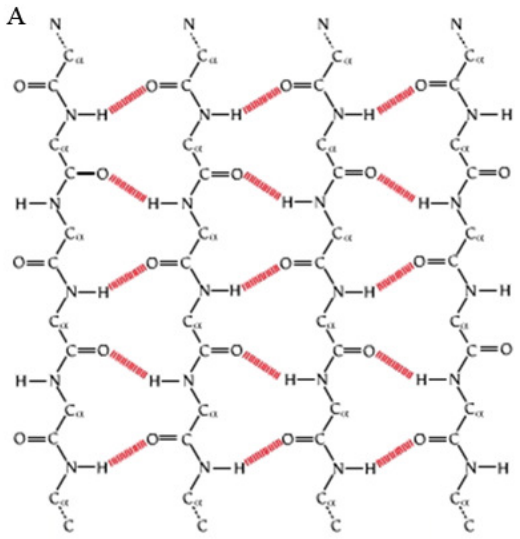

Figure 8. Les Feuillets beta.

Structure du feuillet parallèle, les liaisons hydrogène sont représentées en rouge. Figure adaptée du livre (Branden and Tooze 1998).

Des feuillets mixtes existent mais sont moins fréquents. Les feuillets β sont plissés et présentent généralement une torsion droite. Les chaînes latérales sont positionnées alternativement au dessus puis au dessous du plan du feuillet.

2.2.2.3 Irrégularités au sein des structures répétitives

En pratique, dans les structures protéiques, les structures secondaires répétitives cachent une plus grande complexité que ce que leur description canonique peut laisser imaginer. De nombreuses irrégularités ont été décrites (cf. Figure 9). Ces dernières sont souvent importantes pour la fonction de la protéine.

Ainsi, la plupart des hélices ne sont pas linéaires. En 1998, Barlow et Thornton ont observé que 58% des hélices sont courbées et 17% présentent une cassure (ou *kink*) (Barlow and Thornton 1988). Ces résultats ont depuis été confirmés à plusieurs reprises sur des jeux de données plus récents (Offmann et al. 2007). Par ailleurs, le « π-bulge », aussi appelé α-aneurisme, est une déformation des hélices α. Il est caractérisé par au moins une liaison hydrogène *i*, *i*+5 suivie par un nombre variable de résidus n'établissant pas de liaison hydrogène avec le squelette polypeptidique. Il induit une cassure dans l'axe de l'hélice. Extrêmement peu fréquent, il semble pourtant avoir un rôle fonctionnel important notamment au niveau de sites de liaison (Cartailler and Luecke 2004).

De même, des irrégularités peuvent aussi être observées au niveau des feuillets β. Le « β-bulge » correspond à la formation de liaisons hydrogène entre deux résidus (ou plus) d'un brin et un seul résidu du brin adjacent (Richardson et al. 1978; Chan et al. 1993). Les β-bulges sont fréquents. En moyenne, deux β-bulges peuvent être observés par protéine. Ils interrompent l'alternance des chaînes latérales de part et d'autre du plan du feuillet. Ils accentuent également leur courbure droite. Leur rôle n'est pas encore clairement identifié. Les β-bulges pourraient faciliter les insertions et délétions dans les brins β. Généralement exposés à la surface des protéines, ils jouent également un rôle dans les interactions protéine-protéine.

Par ailleurs, depuis la description des brins β, plusieurs analyses ont montré que des brins pouvaient également être observés seuls à l'extérieur d'un feuillet. Ces brins sont nommés brins E (Eswar et al. 2003). Ils sont clairement distincts des brins β de part des spécificités de séquence particulière (sur-représentation de proline) mais également de part leur exposition au solvant élevée.

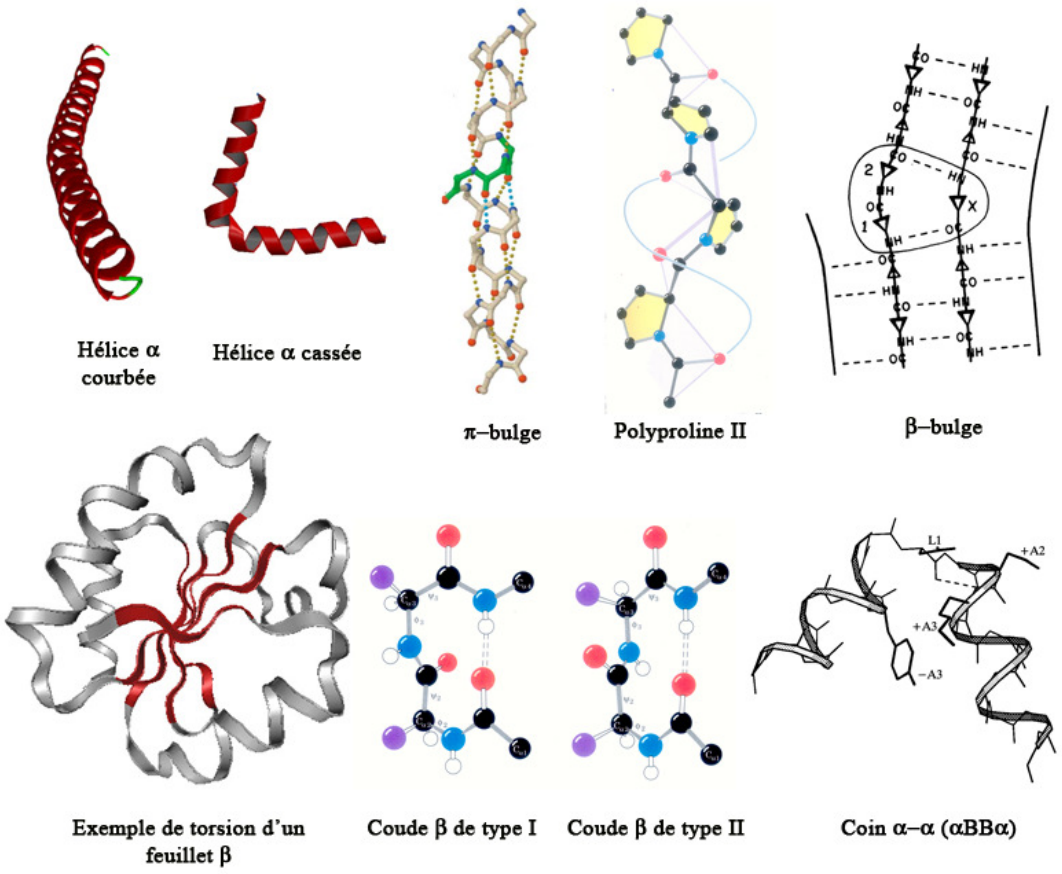

Figure 9. Exemples *(i)* d'irrégularités au sein des structures régulières répétitives et *(ii)* de structures secondaires non-répétitives (2 coudes béta et un coin alpha-alpha).

Hélice α courbées - Exemple extrait d'une protéine potentiellement impliquée dans le transport du phosphate (code PDB 1SUMB (Liu et al. 2005)) et adapté de (Regad et al. 2008). **Hélice α cassée** - Exemple extrait de l'Hémoglobine I de *Lucina pectinata* (code PDB 1B0B (Bolognesi et al. 1999)) selon une attribution des structures secondaires par DSSP (Kabsch and Sander 1983) et adapté de (Benros 2005). **π-bulge** - Les sphères jaunes représentent les liaisons hydrogène de type *i, i*+4 de type hélice α. Les sphères bleus représentent les liaisons hydrogènes *i, i*+5. Figure extraite de (Cartailler and Luecke 2004). **Polyproline II** - Figure extraite de (Voet and Voet 1995). **β-bulge** - Exemple de β-bulge à l'extrémité d'un feuillet β antiparallèle. Les petits triangles représentent les chaînes latérales au dessous du feuillet et les grands représentent les chaînes latérales au dessus. Figure extraite de (Richardson et al. 1978). **Exemple de torsion d'un feuillet β** - Exemple de la Thioredoxine humaine (code PDB 1ERT (Weichsel et al. 1996)) et adapté de (Benros 2005). **Coudes β de types I et II** - Il existe huit types de coudes β différant par leurs angles de torsion. Figure extraite de (Voet and Voet 1995). **Coin α–α de type αBBα** - Les lettres B caractérisent les angles dièdres adoptés par les deux résidus entre les hélices. B est le domaine de la carte de Ramachandran correspondant aux angles des brins β avec un enroulement droit. Figure adaptée de (Wintjens et al. 1996).

Enfin, les polyprolines de type II (PII) sont des hélices très particulières ayant à la fois des caractéristiques d'hélice et de brin β. Elles ont tout d'abord été décrites dans les protéines fibreuses (Cowan et al. 1955). Elles présentent une structure étendue avec un enroulement gauche impliquant 3 résidus par tour. Le pas de l'hélice est de 9,3 Å. Les angles dièdres caractéristiques (Φ, Ψ) sont en moyenne égaux à (-75°, +175°), ces angles sont proches de

ceux observés dans les brins β. Les PII ne sont pas obligatoirement composées d'une succession prolines (Chellgren et al. 2006). Cubellis et associés ont récemment montré qu'elles étaient stabilisées par des interactions non-locales (Cubellis et al. 2005a). Elles pourraient être impliquées dans la formation des fibres amyloïdes et dans les interactions protéine-ADN et protéines-protéines.

2.2.3 Description des structures non-répétitives

Ainsi, les structures secondaires répétitives, constituant le cœur hydrophobe des protéines, ont été largement et précisément étudiées depuis les années 50. Toutefois, elles ne représentent qu'environ 50% des résidus. Les résidus restants sont classiquement considérés comme faisant partie des *boucles* ou en anglais *coils* ou *loops*. Ce sont les structures dites de « connexion » entre hélices et feuillets. Elles sont fréquemment exposées à la surface des protéines et se sont avérées avoir des rôles essentiels dans leurs fonctions, étant par exemple le siège des sites actifs. Dans ces régions, les groupes NH et C'O du squelette polypeptidique forment peu de liaisons hydrogène entre eux. Ils peuvent par contre former des liaisons hydrogène avec les molécules d'eau (Branden and Tooze 1998). Les boucles sont classiquement décrites comme des structures irrégulières sans angles dièdres (Φ, Ψ) spécifiques. Leur longueur peut varier de 2 à une vingtaine de résidus. Cependant, bien que les boucles adoptent des structures très diversifiées, des conformations locales récurrentes ont été observées.

Les *coudes* ou en anglais *turns* sont des boucles courtes impliquant un changement de direction du squelette polypeptidique (voir Figure 9). Ils jouent donc un rôle majeur dans la topologie finale de la protéine. Ils sont constitués de n résidus consécutifs (notés i à $i+n$), la distance entre les Cα des résidus i et $i+n$ devant être inférieure à 7 Å. De plus, les résidus centraux des coudes ne doivent pas être hélicoïdaux afin de ne pas les confondre avec des hélices. Fréquemment, une liaison hydrogène entre le groupe NH du résidu i et le groupe C'O du résidu $i+n-1$ stabilise la structure locale (Venkatachalam 1968). Plusieurs catégories de coudes ont été caractérisées : les coudes γ (n = 3 résidus) (Milner-White 1990), les coudes β (n = 4) (Venkatachalam 1968; Hutchinson and Thornton 1996), les coudes α (n = 5) (Pavone et al. 1996) et les coudes π (n = 6) (Rajashankar and Ramakumar 1996). Au sein de ces catégories, les coudes sont classés en fonction des angles dièdres des résidus centraux. Les coudes β sont les plus fréquents, ils représentent 25 à 30% des résidus (Guruprasad and

Rajkumar 2000; Fuchs and Alix 2005). Une étude récente tend à proposer une nouvelle classification des coudes et présente quelques nouvelles catégories (Koch and Klebe 2009).

Les *boucles Ω* sont des régions de six à seize résidus présentant une distance inférieure à 10 Å entre les résidus aux extrémités et un grand nombre de contacts internes (Leszczynski and Rose 1986). Ces régions sont des boucles globulaires compactes. Elles restent peu explorées.

Les boucles courtes reliant des structures secondaires spécifiques ont également été caractérisées. Les boucles en épingle à cheveux β ou *β-hairpins* liant deux brins β adjacents antiparallèles sont généralement courtes, associées à un retour de la chaîne polypeptidiques de type coude β et sont largement répandues dans les structures protéiques (Sibanda and Thornton 1985). De même, des conformations préférentielles ont été observées pour les coins β–β (β–β corners, entre deux brins β) (Efimov 1991), les épingles à cheveux α, les coins α–α (entre deux hélices α) (Wintjens et al. 1996) (cf. Figure 9), les boucles α–β (entre une hélice α et un brin β consécutifs) et β–α (entre un brin β et une hélice α consécutifs) (Efimov 1993; Wintjens et al. 1998).

L'augmentation du nombre de structures protéiques connues a permis d'établir des classifications des boucles en fonction de leur géométrie. Ainsi, SLOOP créé par Donate *et al.* (1996) puis mis à jour par Burke *et al.* (2000), définit 560 classes de boucles pouvant faire jusqu'à 20 résidus (Donate et al. 1996; Burke et al. 2000). Néanmoins, la grande majorité des classes caractérisent des boucles courtes. Les critères pris en compte pour la classification sont la taille, les structures secondaires répétitives de part et d'autres ainsi que la géométrie des boucles. De même, WLOOP développé parallèlement par Kwasigroch *et al.* puis mis à jour par Wojcik *et al.* contient 183 familles de boucles de 3 à 8 résidus (Kwasigroch et al. 1996; Wojcik et al. 1999). Cette librairie repose sur une classification hiérarchique basée uniquement sur la taille et la similarité de la géométrie des boucles (critère du RMSD[1]). ArchDB fut créée par (Oliva et al. 1997). Son originalité repose sur la prise en compte d'une taille non exacte (taille ± un résidu) pour classer les boucles. Les auteurs prennent ainsi en compte la difficulté à assigner précisément les extrémités des boucles (voir le paragraphe 2.2.5) (Colloc'h et al. 1993; Fourrier et al. 2004). Le type de structures secondaires répétitives

[1] RMSD (Root Mean Square déviation) : distance euclidienne calculée entre les coordonnées de deux fragments superposés de façon optimale. Ce critère permet de savoir si deux fragments possèdent une même géométrie.

bordantes ainsi que la géométrie des boucles (comparaison de domaines d'angles dièdres sur la carte de Ramachandran) sont également considérés. La classification aboutit à 124 sous-classes de boucles en 1997. En 2004, elle est remise à jour et aboutit à 3213 sous-classes comprenant au moins deux boucles (Espadaler et al. 2004). La base de données associée est disponible à l'adresse *http://sbi.imim.es/cgi-bin/archdb/loops.pl*. Cette stratégie de classification a par la suite été utilisée pour classer et caractériser les protéines kinases en fonction de leurs boucles fonctionnelles (Fernandez-Fuentes et al. 2004) (voir Figure 10). D'autres classifications peuvent encore être citées comme celle proposée par Li *et al.* organisant les boucles de 2 à 13 résidus en se basant sur la position dans l'espace des extrémités des structures secondaires répétitives bordantes (Li et al. 1999a; Li et al. 1999b).

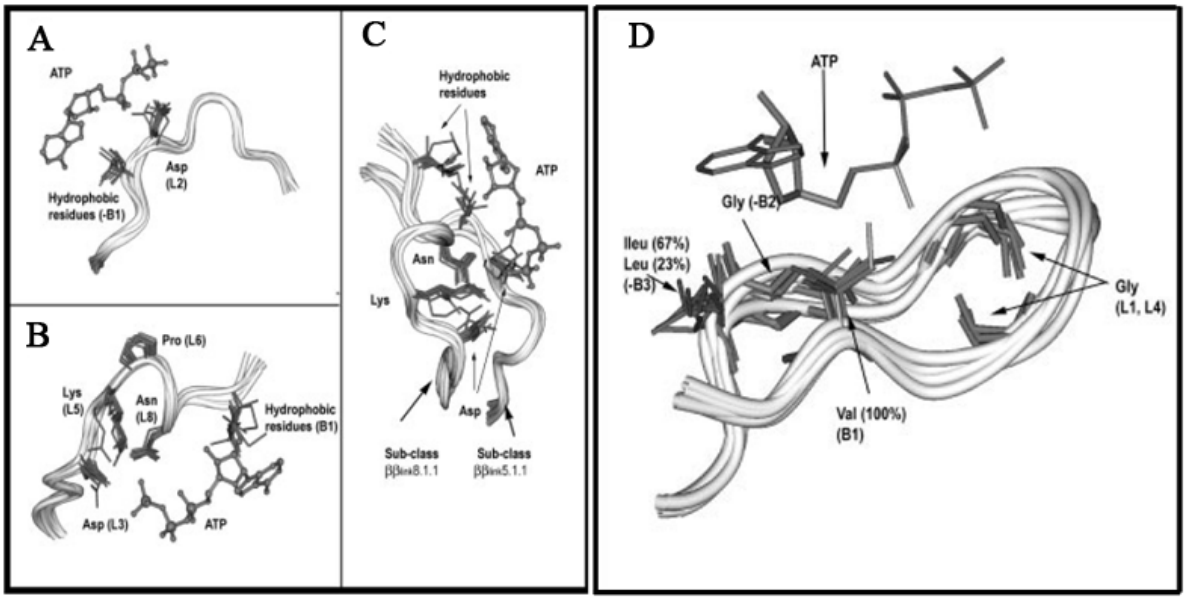

Figure 10. Exemples de sous-classes de boucles obtenues en appliquant la stratégie d'ArchDB au groupe des protéines kinases.

Les noms des boucles présentées correspondent à la classification réalisée par Fernandez-Fuentes et associés (Fernandez-Fuentes et al. 2004). A – Boucle $\beta-\beta_{link}$ 5.1.1 impliquée dans la liaison à l'ATP. B – Boucle $\beta-\beta_{link}$ 8.1.1, caractéristique des boucles catalytiques du repliement des protéines-kinase-like. C – Interactions des boucles $\beta-\beta_{link}$ 5.1.1 et $\beta-\beta_{link}$ 8.1.1 représentant les interactions principales avec l'ATP. D - $\beta-\beta_{hairpin}$ 2.2.2 caractéristique des boucles riches en Glycines observées dans les kinases. Les chaînes latérales des résidus conservés sont représentées en noir. Figure adaptée de (Fernandez-Fuentes et al. 2004).

L'ensemble de ces travaux a contribué à montrer qu'il existait des structures récurrentes au sein des boucles pourtant trop souvent considérées comme non structurées. Toutefois, les régions décrites sont le plus souvent courtes et de longueur moyenne. Les boucles plus longues que 8 résidus restent plus difficiles à caractériser du fait de leur variabilité et de leur faible occurrence.

Par ailleurs, le tableau volontairement détaillé des structures secondaires dressé dans les paragraphes précédents, montre toute la complexité de l'utilisation de cette description. Une caractérisation précise, bien qu'encore incomplète, de la structuration locale des protéines demande une connaissance aigue de ce domaine d'étude. Ainsi, souvent, la description et la prédiction des structures secondaires se limite aux 3 états hélices, feuillets et boucles. Or cette description ne permet pas de caractériser la structure tridimensionnelle des protéines dans son ensemble. L'état boucle, pourtant important pour la fonction (cf. Figure 10), ne caractérise aucune conformation spécifique et les orientations entre éléments de structures répétitives ne sont pas décrites.

2.2.4 Détermination expérimentale des structures secondaires

Le dichroïsme circulaire (DC) est une technique expérimentale qui permet d'obtenir rapidement une estimation de la composition en structures secondaires d'une protéine en solution. Cette méthode requiert seulement environ un microgramme de matériel. Elle repose sur l'interaction entre la lumière polarisée circulaire et les composés optiquement actifs. Deux types de composés actifs existent dans les protéines : les chaînes latérales des résidus aromatiques et les liaisons peptidiques. Dans le domaine des Ultra-violets lointains (180-250 nm), l'environnement des liaisons peptidiques dans les structures secondaires régulières conduit à des spectres caractéristiques (Receveur-Brechot et al. 2006). Par exemple, le spectre des hélices α est positif à 193 nm et négatif à 208 et 222 nm alors que le spectre des feuillets β présente un pic positif à 195 nm et un pic négatif à 218 nm. Les spectres des structures très étendues de type polyproline II sont également identifiables et présentent un pic très négatif à 200 nm (Greenfield 2006).

La spectroscopie infrarouge à transformée de Fourier (IRTF) est basée sur l'absorption d'un rayonnement infrarouge et la détection de vibrations caractéristiques des liaisons peptidiques en fonction de leur conformation. Cette méthode permet de discriminer différents types d'hélices en fonction de leur flexibilité. Les coudes ont également des fréquences de vibration spécifiques (Natalello et al. 2005). Cependant, cette méthode requiert des quantités de protéines plus importantes que le DC. D'autres techniques issues de la spectroscopie Raman existent également. Cependant, ces méthodes ne donnent par d'information résidu-spécifique.

A l'inverse, la Résonance Magnétique Nucléaire (RMN) permet d'accéder à des informations à l'échelle atomique sans nécessiter la résolution de la structure tridimensionnelle de la protéine. La RMN étudie les propriétés magnétiques de noyaux atomiques possédant un spin.

La fréquence de résonance des noyaux reflète en partie leur environnement et donc leur structure secondaire. Ainsi, l'étude des déplacements chimiques par rapport à un niveau de référence permet l'attribution de structures secondaires.

Néanmoins, lorsque les structures tridimensionnelles des protéines (déterminées principalement par cristallographie ou RMN (voir paragraphe 2.3.1) sont disponibles, elles servent naturellement de support pour assigner les structures secondaires de manière résidu-spécifique.

2.2.5 Méthodes d'assignation à partir de la structure 3D (Article 1)

D'un point de vue historique, du fait du nombre limité de structures 3D disponibles, les premières détections de structures répétitives se faisaient manuellement le plus souvent par le cristallographe. Depuis, différentes méthodes d'attribution automatique ont été développées (cf. Tableau 1). Chacune reflète la vision et les problématiques de recherche de ses concepteurs. Elles reposent donc sur des critères variés et conduisent à des résultats qui peuvent ne pas être en accord. De plus, elles ne considèrent pas toutes les mêmes états.

Tableau 1. Méthodes d'assignation des structures secondaires.

Méthodes	Année	Assignation basée sur:
Greer & Levitt	1977	Distance
DSSP	1983	Liaison Hydrogène
DEFINE	1988	Distance
PCURVE	1989	Axe
SSTRUC	1989	Liaison Hydrogène
CONCENSUS	1993	DSSP, DEFINE et PCURVE
STRIDE	1995	Liaison Hydrogène / angle dièdre
PROMOTIF	1996	Liaison Hydrogène / angle dièdre
PSEA	1997	Distance / angle
PROSS	1999	Angle dièdre
XTLSSTR	1999	Distance / angle
DSSPcont	2002	Liaison Hydrogène
SECSTR	2002	Liaison Hydrogène
VORO3D	2004	Voronoï
KAKSI	2005	Distance / angle dièdre
SEGNO	2005	angle / multiple
Beta-Spider	2005	Feuillet β + DSSP pour les hélices α
PALSSE	2005	Cα
Tessellation de Delaunay	2005	Delaunay
SKSP	2007	STRIDE, DSSP, SECSTR, KAKSI, P-SEA, et SEGNO
PROSIGN	2008	Cα

Tableau extrait de (Tyagi et al. 2009a).

Une première classe de méthode d'assignation est basée exclusivement sur les liaisons hydrogène formées entre les groupements chimiques du squelette polypeptidique. DSSP (*Dictionary of Secondary Structure of Proteins* (Kabsch and Sander 1983)) appartenant à cette catégorie, reste aujourd'hui la méthode la plus utilisée par la communauté scientifique. Elle assigne les hélices α, π et 3_{10}, les brins β, les coudes ainsi que des brins dits « isolés ». DSSPcont et SECSTR sont des évolutions directes de DSSP proposant respectivement de prendre en compte la flexibilité de la chaîne polypeptidique (Andersen et al. 2002) (voir paragraphe 6) et d'améliorer la détection des hélices π (Fodje and Al-Karadaghi 2002).

Une seconde catégorie de méthodes repose sur les liaisons hydrogènes mais également sur les angles dièdres. Dans cette catégorie, STRIDE (*secondary STRucture IDEntification method*) est la deuxième méthode la plus communément utilisée (Frishman and Argos 1995). Elle assigne les mêmes états que DSSP. PROMOTIF est également dérivé de DSSP mais ajoute la caractérisation des coudes γ et β, des β-hairpins et des β-bulges (Hutchinson and Thornton 1996). L'assignation des structures secondaires utilisée dans la méthode Pex d'extraction de données de structures protéiques, appartient aussi à cette catégorie (Thomas et al. 2001).

Une troisième catégorie de méthodes encore sont basées sur les distances entre résidus au sein des structures comme DEFINE (Richards and Kundrot 1988) ou dans la méthode de Levitt et Geer (Levitt and Greer 1977). Les angles sont souvent également pris en compte comme dans KAKSI (Martin et al. 2005), PSEA (Labesse et al. 1997) et XTLSSTR (King and Johnson 1999). XTLSSTR assigne notamment les polyprolines de type II.

Les méthodes PROSS et SEGNO définissent une 4^{e} catégorie et reposent exclusivement sur les angles dièdres (Srinivasan and Rose 1999; Cubellis et al. 2005b).

Une cinquième catégorie de méthodes repose sur les coordonnées cartésiennes des Cα (PCURVE (Sklenar et al. 1989), PALSSE (Majumdar et al. 2005) et PROSIGN (Hosseini et al. 2008)). Enfin, des méthodes comme VoTap (*Voronoï Tesselation Assignment Procedure*) utilise une définition des contacts entre résidu grâce à des polyèdre de Voronoï et repose sur une description géométrique des structures secondaires (Dupuis et al. 2005).

Les différences d'attribution de structures secondaires entre méthodes ne sont pas négligeables. En moyenne, le pourcentage de résidus associés aux mêmes états Hélice/Brin/Boucle est seulement de 80%. DSSP et STRIDE sont les plus proches avec 95% d'accord. Par exemple, DSSP peut assigner des hélices longues courbées ou cassées alors que KAKSI assignera plutôt deux hélices droites courtes (Martin et al. 2005). Ces divergences ont des répercussions directes sur la définition des boucles (Fourrier et al. 2004) (voir Figure 11).

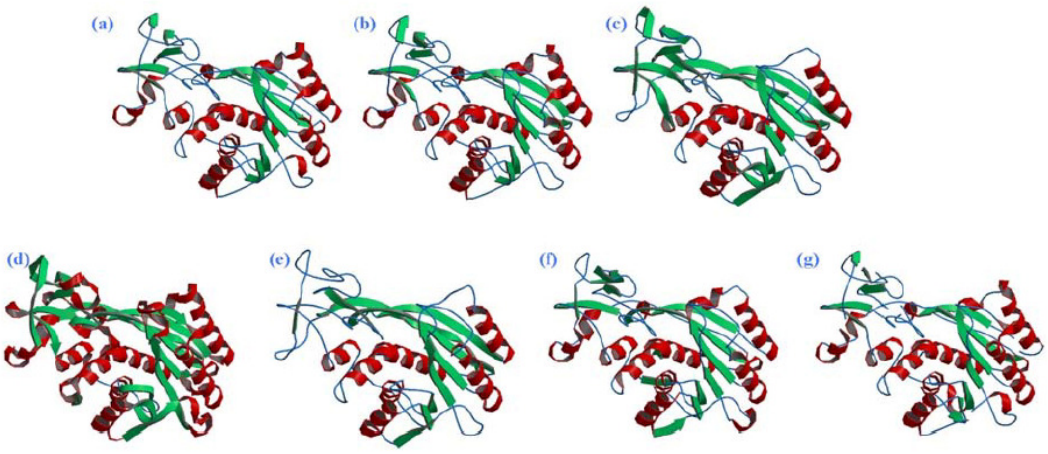

Figure 11. Exemples d'assignations de structures secondaires par différentes méthodes sur la Méthyltransférase Hhai (code PDB 10MH).

Les assignations sont réalisées avec (a) DSSP, (b) STRIDE, (c) PSEA, (d) DEFINE, (e) PCURVE, (f) XTLSSTR and (g) SECSTR. Toutes les assignations ont été réduites à trois états : hélice (red), brin (vert) et boucles (bleu). Figure extraite de (Tyagi et al. 2009b).

Afin de prolonger ces analyses, nous avons mis en évidence dans une récente étude les différences d'assignation des coudes β. Les résultats de huit méthodes d'attribution différentes ont été comparés lors de mon stage de master 2 (Bornot and de Brevern 2006) (Article 1). Dans ce but, nous avons calculé la proportion de résidus associés au même état Hélice/Brin/Coude β/Boucle. Dans la mesure où certaines méthodes n'attribuent pas les coudes, ces derniers ont été assignés *a posteriori* en s'appuyant sur la définition classique du coude β (voir paragraphe 2.1.2.3). Les divergences observées entre les différentes méthodes sont en accord avec les études précédentes. De plus, l'analyse des fréquences de coudes permet de dégager deux groupes de méthodes : STRIDE, DSSP, SEGNO, SECSTR, XTLSSTR et KAKSI assignent environ 20% de coudes alors que PSEA et DEFINE en assignent plus de 25% (Tableau 2 haut). La matrice de confusion présentée dans le Tableau 2 (bas) donne la proportion de résidus assignés en coude β par une méthode également assignés en coude β par une seconde méthode. La matrice est asymétrique du fait de l'attribution différente réalisée par chaque méthode. De manière générale, la confusion entre assignements varie de 64 à 95%. DEFINE est l'exception avec une confusion de seulement 50% vis-à-vis des autres méthodes. STRIDE et DSSP basées toutes deux sur les liaisons hydrogènes donnent les résultats les plus proches mais les autres donnent des résultats clairement distincts. Les coudes β étant assignés après l'attribution des hélices α et des brins β, les divergences entre méthodes concernant les coudes sont largement liées aux divergences concernant les hélices et les brins. Ces difficultés ont des répercussions directes sur l'analyse

de la relation séquence-structure ainsi que sur les méthodes de prédiction de structures secondaires associées, leur évaluation et leur comparaison.

Tableau 2. Distribution des structures secondaires (haut) et Matrice de confusion pour l'assignation des coudes beta (bas).

	α	β	coil[a]	turns[b,c]
DSSP	37.42	21.61	19.78	21.19[b] (20.53)[c]
STRIDE	38.88	22.16	19.06	19.90 (20.24)
XTLSSTR	41.04	19.57	19.57	19.82 (11.39)
PSEA	34.04	24.01	15.70	26.25
DEFINE	28.36	25.92	14.76	30.96
SECSTR	38.74	20.33	21.24	19.69
KAKSI	39.49	22.02	15.53	22.97
SEGNO	35.94	22.41	19.62	22.02

	DSSP	STRIDE	XTLSSTR	PSEA	DEFINE	SECSTR	KAKSI	SEGNO
DSSP	--	89.03	76.39	85.48	59.30	84.28	74.70	87.55
STRIDE	94.49	--	79.33	88.53	59.78	85.12	79.07	91.22
XTLSSTR	81.32	79.55	--	83.76	58.46	74.50	77.22	85.53
PSEA	69.27	67.53	63.81	--	59.75	65.77	73.26	76.16
DEFINE	40.80	39.87	38.71	50.52	--	38.56	46.38	44.51
SECSTR	90.73	86.20	75.17	87.50	60.11	--	77.14	87.03
KAKSI	71.78	72.40	68.53	88.12	59.40	67.68	--	77.40
SEGNO	84.11	82.77	77.30	89.83	59.29	77.64	76.74	--

[a] La fréquence du coil (boucle) correspond aux résidus non associés à des hélices α, des brins β ou des turns (coudes). [b] La fréquence des turns correspond au résidus assignés en tant que β-turns et non associés à une hélice α ou un brin β (notre assignement). [c] Les nombres entre parenthèses correspondent à la fréquence des turns selon l'assignation originale des différentes méthodes. Pour DSSP, elle correspond à la fréquence des états « turns » et « bends ». Tableau extrait de (Bornot and de Brevern 2006).

2.2.6 Relation séquence-structure (Article 2)

Du fait de leur structure chimique, les acides aminés ont des préférences significatives pour certaines structures secondaires (Levitt 1978). Dès 1978, Levitt avait identifié trois principes généraux : *(i)* les résidus aliphatiques favorisent la formation de feuillets β, *(ii)* les résidus avec une chaîne latérale courte ainsi que la proline et la glycine favorisent les coudes, enfin *(iii)* la plupart des autres résidus préfèrent former des hélices. Cette étude se basait sur la première méthode d'assignation automatique des structures secondaires développée par Levitt et Geer (voir paragraphe 2.1.2.5). Depuis, ces observations ont été confirmées et affinées par de nombreuses études (Thomas et al. 2001; Offmann et al. 2007; Malkov et al. 2009).

Les hélices α longues et courtes n'ont pas la même composition en acide aminés (Pal et al. 2003). Les extrémités d'hélices présentent également des particularités. Ainsi, la glycine, très flexible, et la proline, très rigide, sont des résidus « casseurs » d'hélices (Levitt and Greer 1977; Imai and Mitaku 2005). Les extrémités des hélices α sont, de plus, fréquemment stabilisées par des contacts hydrophobes (Aurora and Rose 1998). Les spécificités de séquence au sein des brins β ont également largement été étudiés mais sont plus difficile à saisir car la stabilisation des feuillets est réalisée par des interactions à longues distances (Wouters and Curmi 1995; Offmann et al. 2007).

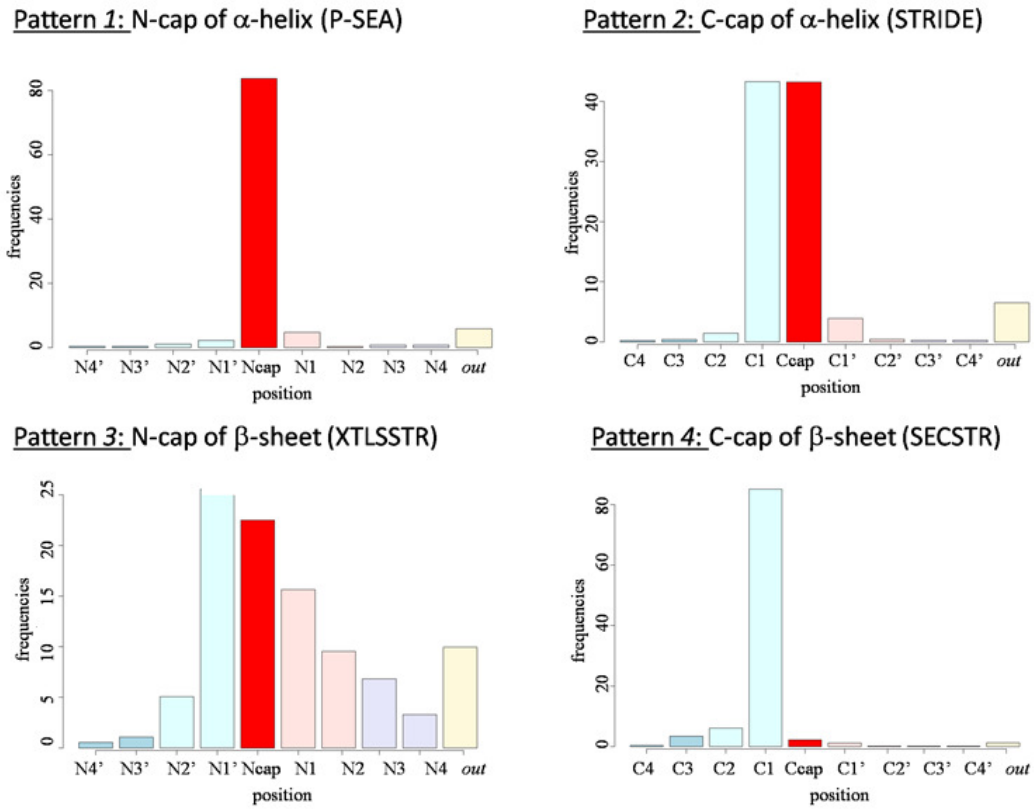

Figure 12. Exemples de divergences entre DSSP et les autres méthodes d'assignation des structures secondaires pour les extrémités des boucles.

En abscisse, sont indiquées les positions dans la séquence. Elles sont centrées sur la position des extrémités N ou C-terminales telles qu'elles sont définies par DSSP. En ordonnées, sont présentées les fréquences auxquelles une méthode assigne l'extrémité de la boucle en une position donnée. Par exemple, pour le *Pattern 2*, STRIDE assigne les extrémités C-terminales des hélices au niveau du même résidu que DSSP dans presque 50 % des cas. Toutefois, pour presque 50 % des cas également, il assigne cette extrémité au résidu précédant celui désigné par DSSP dans la séquence. Les couleurs permettent un repérage visuel des positions. La position des extrémités N ou C-terminales désignées par DSSP est en rouge. Figure extraite de (Tyagi et al. 2009a)

Ces études sont basées sur la méthode d'assignation de prédilection des auteurs. Dans le cadre d'une étude réalisée avec Bernard Offmann et Manoj Tyagi à l'Université de la Réunion, nous avons analysé la relation séquence-structure des extrémités de boucles en tenant compte des

résultats de huit différentes méthodes d'assignements (DSSP, STRIDE, SECSTR, XTLSSTR, PSEA, DEFINE, KAKSI, SEGNO) (Tyagi et al. 2009a) (Article 2). Etonnamment, bien que les extrémités de boucles soient décalées en fonction des méthodes (cf. Figure 12), les spécificités de séquences observées sont très similaires. Elles sont de plus en accord avec les études précédentes. La séquence semble donc autoriser une certaine liberté au niveau de la définition des extrémités des structures secondaires. Cette liberté est peut-être le support physique des adaptations conformationnelles des protéines nécessaires à leur fonction ou à leur stabilité au sein de la cellule (voir section 5).

2.2.7 Méthodes de prédiction à partir de la séquence

La forte relation existant entre la séquence et la structure secondaire adoptée par le squelette polypeptidique a conduit au développement de nombreuses méthodes de prédiction. Elles prédisent en général trois états (Hélice, Feuillet et Boucle) et leur taux de prédiction est noté Q_3.

La première génération de méthodes de prédiction de structures secondaires correspond à des méthodes statistiques. Ainsi, l'une des premières méthodes fut développée par Chou et Fasman. Elle repose sur la propension des acides aminés à adopter chaque conformation (Chou and Fasman 1974b; a). A cette époque, peu de structures tridimensionnelles résolues expérimentalement étaient disponibles. Leurs calculs se basent donc seulement sur 15 protéines en 1974. Le taux de prédiction de ces premières méthodes ne dépassait pas 60%.

Les deux principales améliorations des méthodes de seconde génération a été d'une part de pouvoir prendre en compte plus de données et d'autre part de tenir compte de l'environnement du résidu à prédire dans la séquence (Rost and Sander 1998). L'étude de la relation séquence-structure prend alors en compte les corrélations au sein de fenêtres de séquence. GOR, dont la première version est développée par Garnier, Osguthorpe et Robson, en 1978 est basée sur la théorie de l'information (Garnier et al. 1978). D'autres méthodes prennent en compte diverses informations comme la classe structurale prédite de la protéine (Deleage and Roux 1987) ou la structure de peptides similaires (Levin and Garnier 1988). Ces méthodes obtenaient des taux de prédiction toujours inférieurs à 70%. De plus, les brins β étaient seulement légèrement mieux prédits qu'avec une prédiction aléatoire (Rost and Sander 1998).

Enfin, la troisième génération de méthodes est caractérisée par l'utilisation d'informations évolutionnaires. Pour conserver la fonction des protéines, la pression évolutive a favorisé les mutations ayant peu d'impact sur la structure. La structure est donc classiquement mieux

conservée que la séquence (Illergard et al. 2009). Ainsi, l'utilisation non plus de la séquence cible seule, mais d'alignement de séquences similaires permet notamment de donner plus de poids aux positions conservées. De même, l'utilisation de méthodes d'apprentissage a également fortement contribué à l'augmentation des taux de prédiction. Les méthodes les plus utilisées de nos jours sont PSIPRED (Jones 1999) et SSpro (Pollastri et al. 2002). Elles utilisent toutes deux des réseaux de neurones et sont basées sur une assignation réalisée par DSSP. Des méthodes proposées plus récemment utilisent des chaînes de Markov cachées (HMM) comme (Crooks and Brenner 2004) ou encore des machines à vecteurs support (SVM) comme (Ward et al. 2003). Des méthodes consensus ont également été proposées. Elles permettent généralement une légère augmentation du taux de prédiction (Biou et al. 1988).

La prédiction des structures secondaires apporte des informations structurales importantes. Elle est souvent considérée comme une première étape vers la prédiction des structures tertiaires. Néanmoins, elle n'apporte aucune information quant à l'orientation des différents éléments de structure. La conformation locale des boucles, en particulier, n'est pas ou très pauvrement caractérisée.

2.3 Structure tertiaire

Le troisième niveau d'organisation des protéines est la *structure tertiaire* ou *structure tridimensionnelle globale*. Elle correspond à la forme fonctionnelle de la protéine obtenue après repliement de la chaîne polypeptidique.

Entre les structures secondaires et les structures tertiaires, deux niveaux d'organisation intermédiaires sont classiquement considérés :

- **Les structures super-secondaires.**

Au sein des structures tertiaires, des combinaisons géométriques récurrentes de structures secondaires ont été identifiées. Ces structures dites *super-secondaires* ont fréquemment un rôle fonctionnel important. Ces motifs sont stabilisés par des interactions étroites entre les chaînes latérales des structures secondaires (Efimov 1994). La combinaison Hélice-Coude-Hélice est par exemple impliquée dans la liaison à l'ADN ou la liaison au calcium (voir Figure 13). Elle est donc largement impliquée dans la régulation des activités cellulaires. Le β-hairpin (Figure 10), la clé grecque ou encore le motif β–α–β (Figure 13 B) sont également largement observés dans les protéines, les deux derniers motifs étant notamment impliqués dans la formation de structures en forme de tonneau (voir paragraphe 2.3.3).

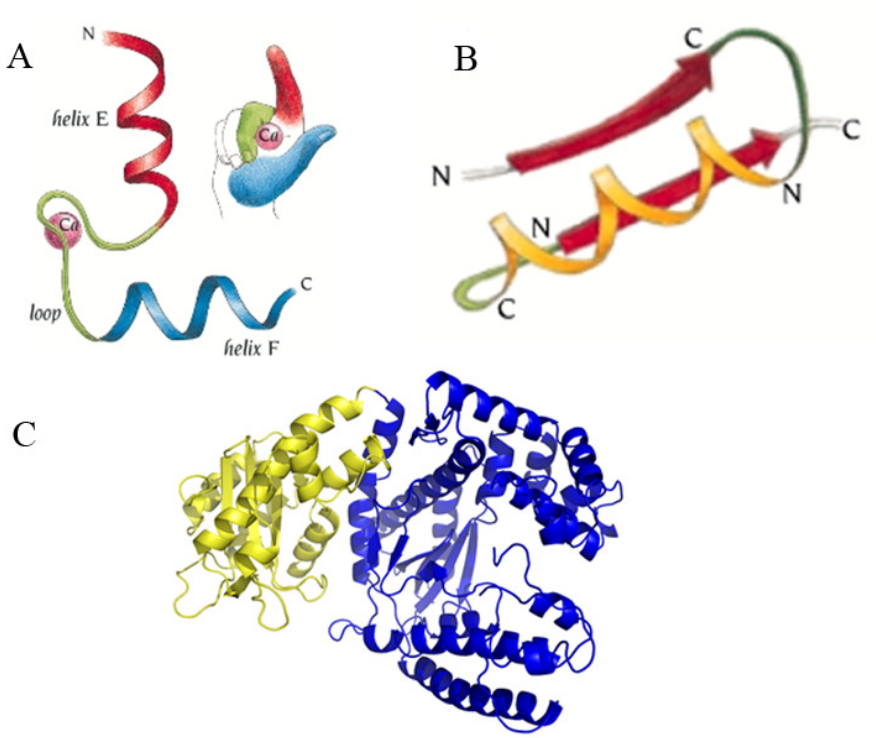

Figure 13. Exemples de structures super-secondaires et d'une protéine organisée en deux domaines.

A – Motif Hélice-Coude-Hélice de liaison au calcium, aussi appelé *EF-Hand*. B – Motif β–α–β. C – Structure de l'ADN polymérase I d'*Escherichia coli*. Sa chaîne polypeptidique se replie en deux domaines distincts représentés en jaune et en bleu (code PDB 2KFZ, (Brautigam et al. 1999)). Les Figures A et B sont extraites de (Branden and Tooze 1998).

- **Les domaines.**

Les motifs de structures secondaires s'organisent en structures globulaires compactes nommées *domaines*. La structure d'une protéine peut-être constituée d'un ou plusieurs domaines (voir Figure 13). Les domaines sont définis comme des unités de repliement et d'évolution indépendantes (voir paragraphes 2.3.3 et 2.3.6). Ils représentent aussi souvent une unité fonctionnelle (Moore et al. 2008). Dans les organismes procaryotes, 40% des protéines seraient multi-domaines. Ce nombre s'élèveraient à 65% chez les eucaryotes (Ekman et al. 2005).

2.3.1 Détermination expérimentale

Il y a 50 ans, en 1958, Kendrew et ses collaborateurs publièrent la première structure tridimensionnelle d'une protéine résolue expérimentalement (Kendrew et al. 1958). Cette structure fut obtenue à partir d'expériences cristallographiques sur la myoglobine. La résolution n'était pas suffisamment élevée pour parvenir à un niveau de détail atomique, néanmoins, ce succès était le début d'une véritable révolution (cf. Figure 14). La première

structure cristallographique haute-résolution fut celle du lysozyme. Elle fut publiée par Phillips et ses collaborateurs en 1965 (Blake et al. 1965). Plusieurs autres modèles haute-résolution suivirent rapidement et les premières analyses liant la structure à la fonction s'imposèrent (Fersht 2008).

Figure 14. Structure de la myoglobine à faible résolution, publiée par Kendrew et collaborateurs en 1958.
La chaîne polypeptidique est blanche et le disque gris représente l'hème. Les graduations sur les échelles en bas des structures mesurent 1 Å. Figure initialement publiée dans (Kendrew et al. 1958) et extraite de (Fersht 2008).

La détermination de structures protéiques tridimensionnelles nécessite d'obtenir un cristal pour chaque protéine étudiée. Dans le but de préserver l'intégrité de la protéine, ce cristal doit contenir 40 à 60% d'eau par volume (Voet and Voet 1995). Cette première étape est particulièrement limitante car la cristallisation dépend de la pureté et de l'homogénéité de la protéine et les cristaux se forment lorsque les molécules précipitent lentement à partir de solution saturée. Certaines protéines présentent des propriétés ne permettant pas l'obtention de cristaux. Le degré de flexibilité de la protéine est notamment un facteur déterminant. Le cristal est ensuite exposé aux rayons X et permet d'obtenir un diagramme de diffraction à partir duquel une carte de densité est calculée. Plusieurs cartes de densité sont ensuite interprétées en termes de coordonnées atomiques puis de modèle tridimensionnel. Les phases de construction et d'affinement du modèle se font principalement aujourd'hui à l'aide d'outils informatiques dédiés calibrés à partir de données empiriques. L'expertise du cristallographe

est également essentielle pour lever les ambiguïtés (Kleywegt and Jones 1997). En générale, une résolution inférieure à 2,5 Å permet l'obtention de modèles corrects.

En 1985, une seconde méthode permettant d'accéder aux structures tridimensionnelles des protéines émergea : Wuthrich et collaborateurs publièrent la première structure résolue par Résonnance Magnétique Nucléaire en solution (RMN) (Williamson et al. 1985). Cette étude ciblait une protéine inhibitrice de protéases. Depuis, la RMN en solution s'est imposée comme une technique complémentaire de la cristallographie, elle est la deuxième plus grande pourvoyeuses de structures protéiques.
Contrairement à la cristallographie, la spectroscopie RMN donne des informations localisées au niveau des atomes de la protéine. Seuls les noyaux des atomes possédant un moment magnétique de spin sont visibles en RMN. Cette technique demande donc le plus souvent un marquage préalable de la molécule avec des isotopes appropriés. Le principe repose sur l'application d'un champ magnétique permettant de synchroniser tous les moments magnétiques de la protéine. Les spécificités de retour à l'équilibre de chaque atome sont ensuite étudiées. Ces dernières dépendent de son environnement dans la protéine et permettent d'obtenir des contraintes et des distances inter-atomiques. Un modèle 3D peut alors être déduit. La RMN est en générale limitée à l'étude de petites protéines. Cependant, elle présente l'avantage de donner des informations sur les protéines en solution, elle est donc une technique de choix pour l'étude de la dynamique des protéines.

D'autres techniques existent. La cryo-microscopie électronique notamment fournit des informations structurales précieuses sur des assemblages protéiques de grande taille et permet ainsi de faire le lien entre molécule et biologie cellulaire (Baumeister and Steven 2000).

2.3.2 La banque de données internationale de structure : la Protein Data Bank (PDB)

La *Protein Data Bank* (PDB) est la base de données internationale de dépôt des structures protéiques. Cette banque fut créée en 1971 et depuis croit de manière exponentielle (Bernstein et al. 1977; Berman et al. 2000). Au début de la cristallographie, certains expérimentateurs ne désiraient pas rendre publique les coordonnées des structures résolues. Il a fallu 30 ans pour que tous les journaux considèrent la déposition des coordonnées atomiques comme un pré-requis à la publication (Fersht 2008). Aujourd'hui, cette banque contient plus de 56 000

structures protéiques (Juin 2009). 89% sont issues de la cristallographie, 14% de la RMN et 0,4% de la cryo-microscopie électronique. Elle est gérée par un consortium international impliquant les USA, l'Europe et le Japon (Berman et al. 2003).

La PDB est extrêmement redondante. Sans filtrage en fonction de l'identité de séquence, cette redondance crée des biais dans les analyses statistiques. Il convient donc de la prendre en compte.

En raison de difficultés techniques et du coût humain élevé nécessaire à la résolution expérimentale des structures, le fossé entre le nombre de séquences protéiques connues (8,7 millions de séquences, voir paragraphe 2.1.2) et le nombre de structures continue de se creuser. Des consortiums de génomique structurale tentent de planifier et d'organiser les efforts au niveau international pour choisir des cibles permettant une augmentation significative des connaissances mais également pour réduire les coûts (Chandonia and Brenner 2006). Dans ce cadre, les projets PSI (*Protein Structure Initiative*), RSGI (*Riken Structural Genomics/Proteomics Initiative*) et SPINE (*Structural Proteomics in Europe*) furent lancés respectivement en 2000 au Etats-Unis, en 2001 au japon, et en 2002 en Europe.

2.3.3 Classification des structures protéiques en repliements connus

Dans le but d'organiser les connaissances, des classifications de domaines protéiques en fonction de leur structure ou *repliement* se sont développées. SCOP (*Structural Classification Of Proteins*) et CATH (*Class, Architecture, Topology, Homologous superfamily*) sont les deux classifications les plus utilisées. Toutes deux organisent les repliements de façon hiérarchique. Elles ont été développées avec l'intention de mettre en évidence les relations phylogénétiques entre protéines. Elles reposent sur l'observation selon laquelle la structure est généralement mieux conservée que la séquence du fait des pressions évolutives exercées pour le maintien de la fonction (Thornton et al. 1999).

SCOP classe les repliements selon quatre niveaux hiérarchiques (Murzin et al. 1995):

- Les *Familles* au sein desquelles les protéines partagent une identité de séquence de plus de 30% ou une identité plus faible mais une similarité de structure et de fonction très importante. Au sein de ces familles, les protéines partagent une origine évolutionnaire commune.
- Les *Super-familles* regroupent des familles présentant de faibles similarités de séquence mais pour lesquelles des similarités de structure et de fonction suggèrent une origine évolutionnaire commune.

- Les *Repliements* regroupent des super-familles présentant des structures secondaires similaires et connectées selon la même topologie. A ce niveau, aucune origine évolutionnaire commune n'est plus garantie.
- Les *Classes* rassemblent les repliements présentant une composition similaire en structures secondaires. Quatre principales Classes sont classiquement identifiées : les domaines dits *Tout-Alpha* (constitués majoritairement d'hélice α), *Tout-Beta* (majorité de brins β), *Alpha/Beta* (notés α/β, alternance d'hélices et de brins) et *Alpha + Beta* (notés $\alpha+\beta$, constitués de régions distinctes avec une majorité d'hélices d'une part et de brins d'autre part).

Grâce à une inspection manuelle de toutes les structures protéiques classées, SCOP propose une classification de très grande qualité et est très vite devenue une référence pour la communauté scientifique internationale. Cette place importante laissée à l'expertise humaine est encore d'actualité aujourd'hui malgré une légère automatisation rendue nécessaire par l'augmentation importante du nombre de structures disponibles (Andreeva et al. 2008).

CATH est également une classification hiérarchique (Orengo et al. 1997). Elle est construite selon une philosophie similaire à celle de SCOP et organise les domaines en fonction de leurs similitudes de séquence puis de structure. Cependant, à la différence de SCOP, sa construction est majoritairement automatisée. Cinq niveaux sont définis (Figure 15) : *(i)* la *classe* caractérise la composition en structures secondaires et leur arrangement, *(ii)* l'*architecture* décrit plus précisément l'arrangement des structures secondaires sans tenir compte de leur connectivité, *(iii)* la *topologie* caractérise la structure générale du cœur du domaine et la connectivité des éléments de structures secondaires, *(iv)* le niveau des superfamilles homologues regroupe les domaines pouvant partager un ancêtre commun en fonction de similarités de séquence et /ou de structure, enfin, *(v)* le dernier niveau des familles de séquences regroupe les domaines présentant des similarités de séquences significatives.

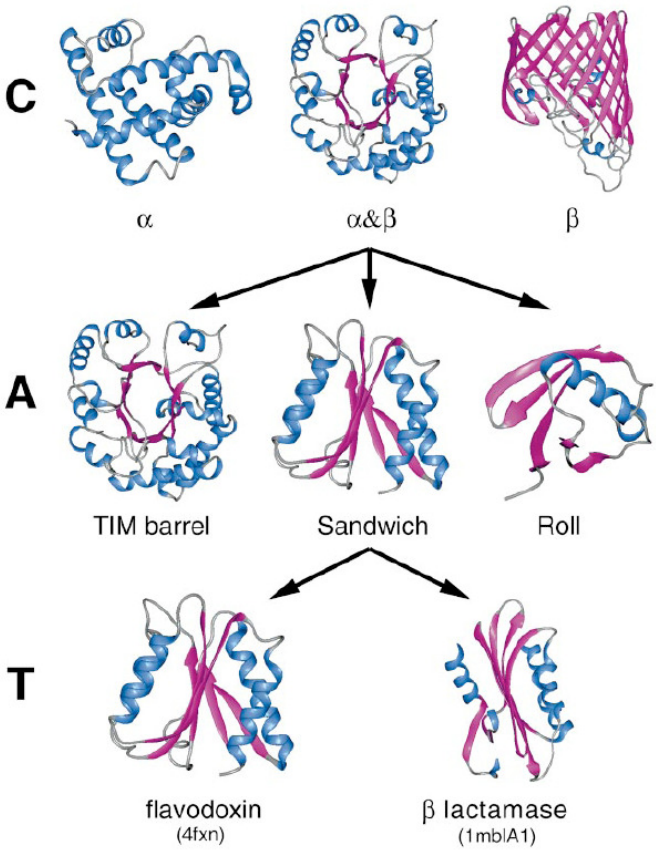

Figure 15. Représentation des trois premiers niveaux Classe, Architecture, Topologie de la classification CATH.

Les hélices sont représentées en bleus, les feuillets en magenta et les boucles en gris. Figure extraite de (Orengo et al. 1997).

Ces classifications ont eu un impact très important sur la compréhension des processus évolutifs au sein des structures protéiques et des relations séquences-structure-fonction. Cependant, il convient de rester prudent face au caractère automatique de CATH et surtout face à la classification de protéines de fonction inconnue issues des projets de génomique structurale. En effet, les noms donnés aux différents sous-groupes de CATH ou SCOP peuvent sous-entendre l'association à une fonction précise. CATH propose par exemple des topologies du nom de *Thiol Ester Dehydrase* ou *Methane Monooxygenase Hydroxylase*. Or, l'assertion selon laquelle une similarité de structure implique une similarité de fonction n'est pas toujours vérifiée et ne permet donc pas une assignation fonctionnelle directe (Doppelt-Azeroual 2009). La structure reste cependant très informative et permet de proposer des hypothèses qui devront être vérifiées expérimentalement.

2.3.4 Forces de repliement et contacts maintenant les structures protéiques

Le repliement d'une protéine et la stabilisation de sa structure dépendent des interactions non covalentes établies entre les résidus de la chaîne polypeptidique. Ces interactions sont de différents types :

- Les interactions électrostatiques comprennent les liaisons hydrogènes et les liaisons ioniques (ou pont salins). Les liaisons hydrogènes participent largement à la stabilisation des structures protéiques. Elles peuvent se former entre les chaînes latérales, dans le squelette polypeptidique ou avec l'eau. Elles stabilisent notamment la formation des structures secondaires (voir paragraphe 2.2). Les liaisons ioniques impliquent des groupements chargés de façon opposée comme par exemple le groupement basique de la lysine ou de l'arginine et le groupement acide du glutamate ou de l'aspartate (Mitchell et al. 1992). Des interactions particulières peuvent également s'établir avec les résidus aromatiques par délocalisation des électrons de leur noyau (Chakrabarti and Bhattacharyya 2007).
- Les interactions de van der Waals correspondent aux forces d'attraction et de répulsion existant entre les atomes et ont principalement un rôle à courte distance.
- L'effet hydrophobe est le facteur le plus important pour la stabilité d'une structure protéique et une force majeure dans le processus de repliement. L'effet hydrophobe est principalement entropique. Les résidus hydrophobes ne peuvent former de liaison hydrogène avec les molécules d'eau et obligent celles-ci à s'organiser localement en cages de solvatation. L'enfouissement d'une grande partie des résidus hydrophobes au sein de la protéine réduit la surface hydrophobe permet un gain d'entropie de l'ensemble des molécules d'eau et la protéine (Rose and Wolfenden 1993).

A ces liaisons non covalentes, s'ajoutent les ponts disulfures entre cystéines (voir paragraphe 2.1.1). Ces liaisons covalentes sont impliquées dans la stabilisation de certaines protéines.

Selon Anfinsen, "toute l'information nécessaire pour obtenir la conformation native d'une protéine dans un environnement donné est contenue dans l'enchaînement des acides aminés" (Anfinsen 1973). Il défend ainsi l'hypothèse dite « thermodynamique » selon laquelle les interactions déterminées par la séquence mènent à la conformation de plus basse énergie. Néanmoins, l'exploration aléatoire de toutes les conformations possibles pour trouver la structure de plus basse énergie n'est pas réalisable dans des temps raisonnables. Il est donc nécessaire de faire l'hypothèse d'un nombre limité de chemins de repliement possibles. Ainsi,

Levinthal propose la formation simultanée de petits noyaux structurés dans plusieurs régions d'une même chaîne polypeptidique qui initieraient et accélèreraient le repliement (Levinthal 1968; Benhabilès et al. 2000). Ces deux hypothèses thermodynamique et cinétique sont complémentaires et coexistent dans les deux modèles émergeant principalement aujourd'hui : le modèle de *diffusion-collision* et le modèle de *nucléation-condensation* (voir Figure 16).

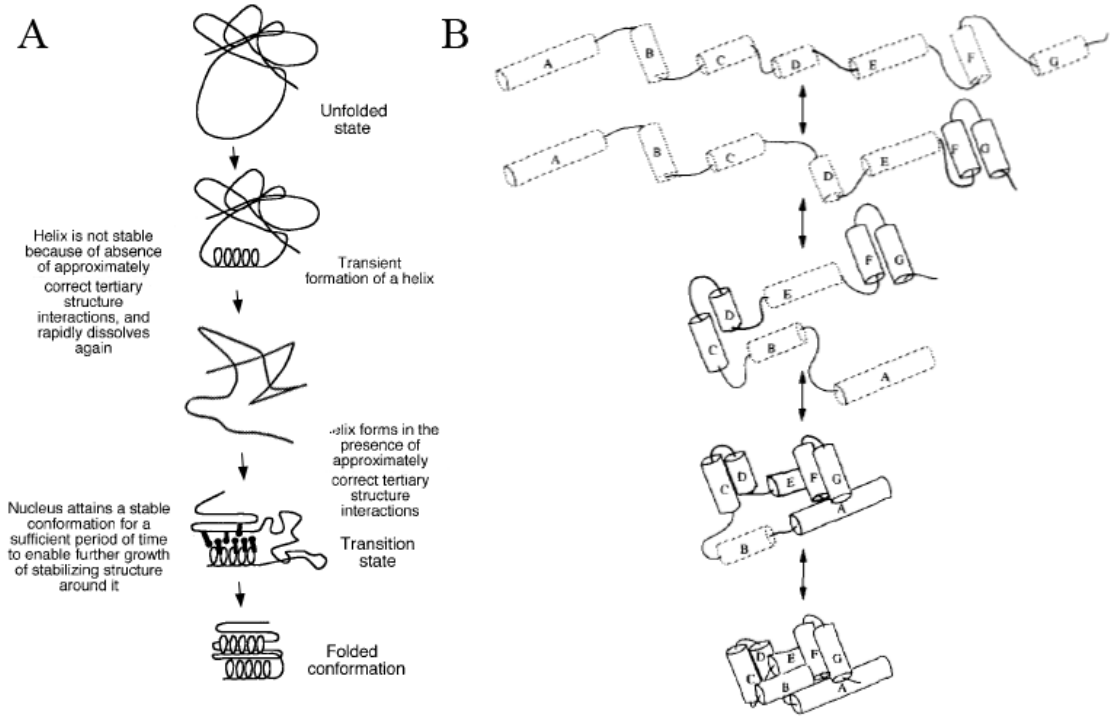

Figure 16. Modèles de repliement.

A – Exemple d'une séquence d'évènements possibles selon le Modèle de *nucléation-condensation* : une hélice se forme de façon transitoire au sein de l'état déplié pendant que la structure tertiaire se replie et se déplie autour d'elle. Après l'échantillonnage d'un grand nombre de conformations possibles, une conformation proche de la structure native se forme autour de l'hélice naissante et la stabilise. Cette stabilisation de certaines interactions secondaires et tertiaires (nucléation) favorise la suite du repliement (condensation). Figure adaptée de (Nolting and Andert 2000). B – Séquence d'évènements illustrant le modèle de *diffusion-collision*. L'état de départ est un ensemble de conformations aléatoire Les *microdomaines* (A-G) sont individuellement instables et forment des structures secondaire de façon transitoire (leur caractère transitoire est indiqué par les pointillés). Ces structures secondaire sont celles ou proches de celles présentent dans l'état natif. La rencontre de ces *microdomaines* transitoirement repliés par diffusion mène à la formation d'intermédiaires stables (représenté par une ligne continue) grâce à des interactions hydrophobes. Ces *multi-microdomaines* intermédiaires interagissent ensuite ensemble pour former une structure tridimensionnelle stable. Figure extraite de (Karplus and Weaver 1994).

Ces deux modèles proposent des chemins de repliement différents mais ne s'excluent pas entre eux. Certaines protéines comme l'inhibiteur 2 de la chymotrypsine semblent suivre le modèle de *nucléation-condensation*, selon lequel la formation de noyaux de repliement catalyse le repliement en une structure tertiaire. D'autres protéines comme la barnase ou l'apomyoglobine semblent se replier selon le modèle de *diffusion-collision* proposé par

Karplus (Karplus and Weaver 1976; 1994), *i.e.* des noyaux de repliement se forment et sont stabilisés de façon concomitante à la formation de la structure tertiaire (Nolting and Andert 2000; Travaglini-Allocatelli et al. 2009). Ainsi, tout semble se passer comme si les protéines résolvaient leur problème d'optimisation globale du repliement par la résolution d'une série d'optimisations locales (Dill et al. 2008).

2.3.5 Interactions préférentielles entre acides aminés au sein des structures (Article 3)

L'étude des contacts entre résidus est une technique classique d'exploration des structures. Elle permet prédire leur vitesse et leur mécanisme de repliement (Plaxco et al. 1998; Gromiha and Selvaraj 1999; Li et al. 2008; Gromiha 2009), de développer des potentiels d'énergie empiriques (Miyazawa and Jernigan 1996; Zhang and Skolnick 1998), ou encore d'identifier des regroupements de résidus jouant un rôle fonctionnel ou structural essentiel (Dosztanyi et al. 1997; del Sol and Carbonell 2007). La prédiction des contacts entre résidus à partir de la séquence est également un champ de recherche actif mais reste encore actuellement un défi (Izarzugaza et al. 2007).

La définition d'un contact n'est en fait pas triviale. De façon classique, deux résidus sont considérés en contact s'ils sont situés à une distance inférieure à un seuil τ l'un de l'autre (cf. Figure 17). Selon les études, τ et les atomes pris en compte pour calculer les distances peuvent être variables. De plus, il est possible de considérer uniquement les contacts à courte, moyenne ou longue distance dans la séquence. Par exemple, dans le cadre d'analyses de la relation entre les contacts au sein de structures 3D et la vitesse de repliement de ces dernières, Plaxco *et al.* considèrent que de résidus sont en contact si leur atomes lourds sont à moins de 6 Å (Plaxco et al. 1998). De leur côté, Gromiha et Selvaraj ne prennent en compte que les Cα et un seuil τ égale à 8 Å (Gromiha and Selvaraj 2001). Dans le cadre de la prédiction des contacts, τ vaut le plus souvent 8 Å mais peut prendre des valeurs allant jusqu'à 12 Å (Pollastri and Baldi 2002; Vullo et al. 2006). A l'inverse, les études plus fines portant sur les interactions préférentielles entre résidus se concentrent sur les atomes lourds des chaines latérales et prennent en compte des distances de contacts de 5,5 Å (Thomas et al. 2002) voire 4 Å (Samanta et al. 2000).

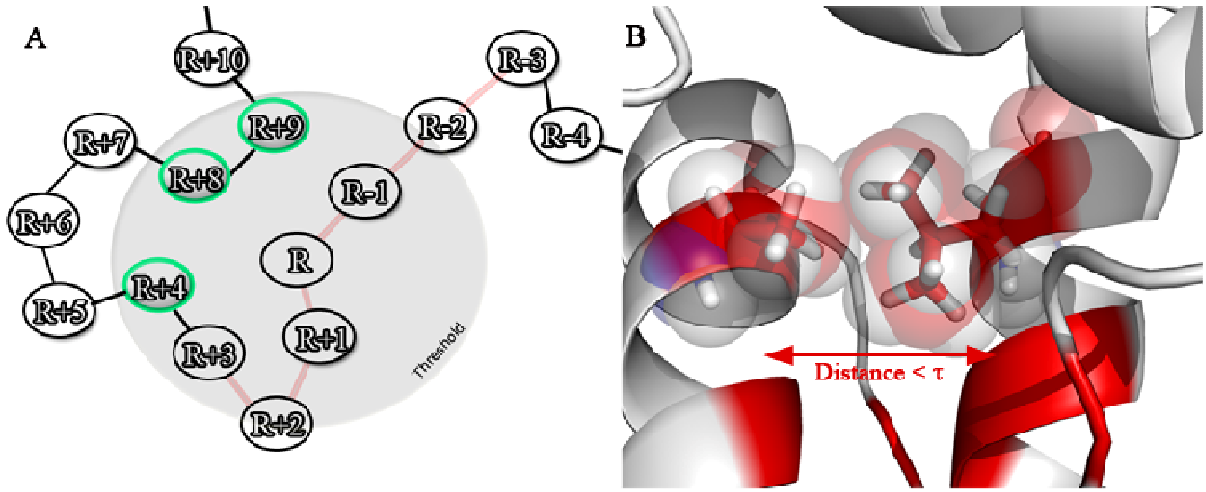

Figure 17. Définition des contacts entre résidus.

A - Le résidu d'intérêt est noté R. Le cercle gris représente la distance τ maximale permettant de définir l'existence d'un contact. Les résidus enchainés en rouge sont les résidus voisins du résidu d'intérêt dans la séquence et peuvent ne pas être pris en compte pour ne pas considérer les interactions à courte distance. Ici, les six résidus voisins ne sont pas pris en compte. Ainsi, seuls les résidus verts sont en contact avec R. Figure extraite de (Faure et al. 2008). B – Exemple d'un contact entre une alanine et une valine au sein du domaine de liaison à l'ADN de la protéine Ku70 (code PDB 1JJR). La distance calculée dépend des atomes considérés (voir ci-dessous).

Lors du stage de master 1 de Guilhem Faure (stage que j'ai encadré), nous avons réalisé une étude systématique permettant de comparer l'impact de ces différentes définitions sur les contacts préférentiels observés entre résidus (Faure et al. 2008) (Article 3). Ces analyses ont de plus été étendues pour mettre en évidence des particularités en fonction de la longueur des protéines, de leur classe SCOP, des structures secondaires ou encore de l'accessibilité au solvant des résidus. Ainsi, cinq distances ont été analysées :

- La distance entre les Cα des résidus, notée Cα.
- La distance entre les Cβ, notée Cβ. (Le Cα est considéré dans le cas de la glycine).
- La distance minimale entre les atomes lourds du squelette polypeptidique correspondant à deux résidus, notée BB.
- La distance minimale entre les atomes lourds des chaînes latérales des résidus, notée SC.
- La distance minimale entre tous les atomes lourds des résidus, notée ALL.

De plus, des valeurs de τ allant de 4 à 20 Å ont été étudiées. Pour ne pas prendre en compte les interactions au sein des structures secondaires, les interactions à courte distance dans la séquence ont été écartées. Ainsi, les six résidus entourant le résidu d'intérêt dans la séquence ne sont jamais considérés en contact, conformément à différentes études précédentes (Punta

and Rost 2005; Gelly et al. 2006a). Ces analyses ont été réalisées sur une banque de structures non redondante représentative de l'espace des structures protéiques.

Nous avons tout d'abord montré que la distribution du nombre de contacts moyen est clairement dépendante des atomes pris en compte pour le calcul de la distance inter-résidus (voir Figure 18). Globalement, les distances Cα et Cβ donnent des résultats similaires. De même, les distances BB et SC donnent également des résultats très proches. Cependant, pour τ = 8 Å, les dernières définissent deux fois plus de contacts que les premières. De même, la distance ALL indique deux fois plus de contacts que BB et SC et quatre fois plus que Cα et Cβ. De plus, l'un des résultats les plus intéressants est que les contacts définis par une distance ne sont pas forcément pris en compte par une autre distance. Par exemple, seuls 58% des contacts définis par $ALL^{\tau=4}$ sont aussi définis par $C\alpha^{\tau=8}$ et réciproquement, 44% des contacts définis par $C\alpha^{\tau=8}$, ne le sont pas par $ALL^{\tau=4}$. De même, seuls 22% des contacts définis par $SC^{\tau=4}$ sont aussi définis par $C\alpha^{\tau=8}$. Les interactions étudiées dans le cadre des méthodes de prédiction des contacts et celles étudiées dans le cadre de l'analyse fine des contacts entre résidus sont donc différentes. De plus, la prédiction prend en compte plus de contacts.

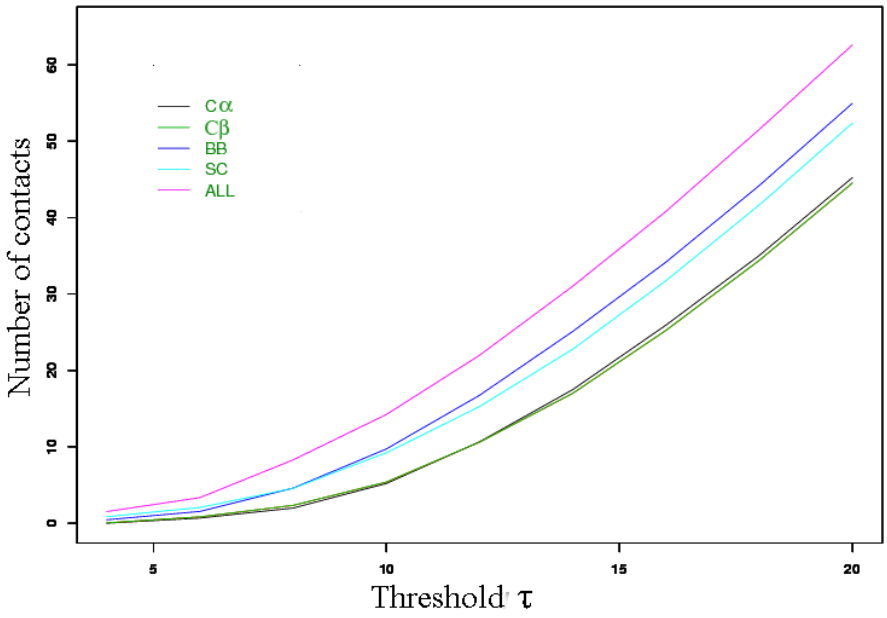

Figure 18. Evolution du nombre de contacts moyen par résidu.

Le nombre de contacts observés est présenté en fonction du seuil τ utilisé pour définir un contact. Les résultats proposés correspondent aux distances Cα, Cβ, BB, SC et ALL. Figure extraite de (Faure et al. 2008).

La distribution des interactions préférentielles avec $C\alpha^{\tau=8}$ montre comme attendu, des interactions privilégiées entre les hydrophobes et entre les aromatiques. Comme exposé dans

le paragraphe 2.3.4, le regroupement des résidus hydrophobes dans le cœur des protéines est largement favorisé énergétiquement. Par ailleurs, les résidus aromatiques, grâce à leur cycle, peuvent interagirent entre eux de multiples manières, *e.g.,* interactions entre les noyaux électroniques (π–π) ou interactions avec un donneur *X* de proton (*X*-H…π). Ils participent ainsi à la stabilisation de la structure tertiaire (Samanta et al. 2000; Thomas et al. 2002). Ils sont notamment impliqués dans le maintien des feuillets β ou dans la stabilisation des interactions brin β / hélice α ou brin β / boucle. De plus, une forte préférence des cystéines à interagir avec d'autres cystéines a été confirmée. Environ un quart des cystéines sont impliquées dans des ponts disulfures participant également à la stabilisation de la structure tertiaire. Ces résultats sont similaires à ceux obtenus avec les distances $C\alpha^{\tau=i}$ pour *i* allant de 6 à 20. En revanche, les deux tiers des interactions préférentielles mises en évidence avec $C\alpha^{\tau=8}$ et $SC^{\tau=4}$ diffèrent de façon significative. Les interactions histidine-histidine, tryptophane-tryptophane et cystéine-cystéine sont notamment renforcées avec $SC^{\tau=4}$. De plus, 18 des 20 acides aminés voient diminuer leur propension à entrer en contact avec la glycine. Ceci est du aux plus fortes contraintes imposées par la distance $SC^{\tau=4}$ combinées à l'absence de chaîne latérale pour la glycine, *i.e.* pour enregistrer un contact, la chaîne latérale d'un résidu donné doit véritablement pointer vers le Cα de la glycine. Des spécificités ont également été mises en évidence en fonction de la distance dans la séquence mais aucun changement drastique. Pour les protéines de plus de 150 résidus, la taille de la chaîne polypeptidique ne semble pas avoir d'impact sur les interactions préférentielles. En revanche, les petites protéines présentent de fortes spécificités de contacts probablement due à une composition en acides aminés légèrement différente. La préférence de contact entre cystéines et entre tryptophanes est encore renforcée. En revanche, les classes SCOP, pourtant associées à des distributions nettement spécifiques d'acides aminés, semblent n'avoir aucun impact sur les interactions préférentielles.

Une attention particulière a parallèlement été portée sur les résultats des méthodes de prédiction du positionnement des chaînes latérales du point de vue des contacts. Le nombre relativement restreint de structures protéiques disponibles renforce la pertinence des techniques bioinformatiques de modélisation comme la modélisation pas homologie, les techniques d'enfilage (*threading*) ou les approches *ab initio* ou *de novo* (voir paragraphe 2.3.8). Ces différentes techniques permettent de prédire une conformation du squelette polypeptidique. Le positionnement des chaînes latérales des résidus est ensuite réalisé dans un second temps et reste encore un problème d'optimisation difficile. Ainsi, des méthodes de

prédiction de la conformation des chaînes latérales ont été développées. SCWRL est la méthode la plus utilisée aujourd'hui (Dunbrack and Karplus 1993; Bower et al. 1997; Dunbrack and Cohen 1997). Elle est basée sur une fonction de score et sur une librairie de rotamères dépendante de la conformation du squelette polypeptidique. La prédiction repose sur la théorie des graphes permettant de réduire la combinatoire des positions possibles (Canutescu et al. 2003). Respectivement 82,6% et 73,7% des angles dièdres χ^1 and χ^{1+2} prédits sont correctes, *i.e.* à ± 40% des angles réels[2]. La méthode SCATD est basée sur SCWRL et utilise une optimisation de l'algorithme de recherche permettant une accélération du traitement (Xu 2005). D'autres méthodes existent. SCCOMP repose principalement sur une fonction de score prenant en compte la géométrie des chaînes, leurs propriétés chimiques, et leur surface accessible (Eyal et al. 2004). SCAP utilise une librairie de rotamères prenant en compte quatre angles dièdres et indépendante de la géométrie du squelette polypeptidique (Xiang and Honig 2001). Son processus de minimisation repose sur le champ de force CHARMM. Enfin, IRECS utilise les probabilités d'apparition des différents rotamères. Dans un second temps, un processus itératif basé sur une fonction d'énergie permet de sélectionner les meilleurs candidats.

De façon classique, il est souvent considéré que ces méthodes donnent des résultats similaires. Dans notre étude, nous avons mis en évidence de fortes divergences entre ces différentes méthodes du point de vue des contacts entre résidus résultant du positionnement des chaînes latérales (Faure et al. 2008). Seuls 75% des contacts prédits par SCWRL sont par exemple prédits par IRECS (Tableau 3). Réciproquement, seuls 70% des contacts trouvés par IRECS, le sont également par SCWRL. D'autre part, la distribution des contacts prédits est très divergente de celle observée dans la banque de données de structures cristallographiques. Avec la distance $SC^{\tau=4}$, le nombre de contacts prédits par la plupart des méthodes est plus faible que dans la banque de données. A l'inverse, SCAP génère 53% de contacts supplémentaires. Par ailleurs, seuls 60% des contacts observés dans la banque de données sont prédits par SCWRL. Ce pourcentage varie entre 55 et 64 % pour les autres méthodes. SCWRL par exemple, génère une sur-représentation de ponts disulfures et une sous-représentation des interactions entre résidus chargés.

[2] L'angle χ^1 correspond à l'angle dièdre d'une chaîne latérale le plus proche de son point d'attache sur le squelette polypeptidique. Ainsi, la notation χ^{1+2} se réfère aux deux premiers angles dièdres d'une chaîne latérale.

Tableau 3. Analyse des contacts prédits par les méthodes de positionnement des chaînes latérales (distance $SC^{\tau=4}$).

	Protein (%)	Contact numbers (%)	DB (%)	SCWRL (%)	SCATD (%)	IRECS (%)	SCAP (%)	SCCOMP (%)
DB	100.0	–	–	60.5	54.8	64.0	59.6	61.2
SCWRL	100.0	−12.2	68.9	–	62.0	75.7	64.4	71.5
SCATD	100.0	−5.8	64.5	64.1	–	68.3	61.1	71.1
IRECS	100.0	−15.0	68.0	70.6	61.7	–	61.4	67.9
SCAP	99.4	+53.1	36.9	36.7	33.9	37.5	–	35.5
SCCOMP	98.9	−1.4	61.2	62.8	60.4	64.0	55.1	–

Tableau extrait de (Faure et al. 2008).

Cette étude montre l'intérêt de prendre en compte les contacts entre résidus pour évaluer les méthodes de prédiction de la conformation des chaînes latérales. Nous avons mis en évidence différents biais qui pourraient avoir un fort impact sur les approches de modélisation des structures protéiques. Une correction des prédictions pourrait être possible en se basant sur les contacts préférentiels observés au sein des structures cristallographiques. De même, une combinaison de ces différentes méthodes permettrait peut-être une amélioration des résultats.

2.3.6 Caractérisation de sous-unités protéiques compactes (Article 4)

Les structures protéiques sont constituées d'un ou plusieurs domaines. Classiquement, chaque domaine est vu comme une unité de repliement et de fonctionnement (Moore et al. 2008). Les méthodes d'analyse phylogénétique et les approches de modélisation moléculaire donnent généralement de meilleurs résultats sur les domaines (Ponting and Russell 2002).

Ainsi, le découpage des structures protéiques en domaines est un outil largement utilisé pour la classification des repliements et pour l'annotation fonctionnelle (Holland et al. 2006; Joshi 2007 ; Taylor 2007). Les méthodes automatiques de découpage des structures en domaines sont généralement basées sur l'hypothèse selon laquelle les interactions au sein des domaines sont plus fortes que les interactions entre les domaines. La majorité des approches maximisent le nombre de contacts au sein des domaines (PUU (Holm and Sander 1994), DOMAK (Siddiqui and Barton 1995) and 3Dee (Dengler et al. 2001; Siddiqui et al. 2001), DETECTIVE (Swindells 1995), DALI (Holm and Sander 1998), STRUDL (Wernisch et al. 1999), DomainParser (Xu et al. 2000; Guo et al. 2003), Protein Domain Parser (Alexandrov and Shindyalov 2003) et DDOMAIN (Zhou et al. 2007)). Des méthodes alternatives proposent un découpage hiérarchique de la protéine en unités compactes (Wetlaufer 1973; Lesk and Rose 1981; Wetlaufer 1981; Wodak and Janin 1981; Sowdhamini and Blundell 1995; Tsai and Nussinov 1997; Pugalenthi et al. 2005). Ces unités compactes pourraient correspondre à des noyaux de repliement apparaissant lors du processus de formation de la

structure tridimensionnelle native. La méthode du *Protein Peeling* développée au sein du laboratoire repose sur ce concept (Gelly et al. 2006a; Gelly et al. 2006b). Elle permet d'identifier des *Unités Protéiques* (UP) tout au long de la séquence, en maximisant le nombre de contacts au sein des unités et en minimisant le nombre de contacts entre elles. Une UP est donc une sous-région compacte de la structure tertiaire correspondant à un fragment de séquence (cf. Figure 19). Pour une protéine donnée, une matrice de contacts est calculée en se basant sur une distance $C\alpha^{\tau=8}$. Celle-ci est ensuite transformée en probabilité de contact grâce une fonction logistique. Un algorithme basé sur le coefficient de corrélation de Matthews entre les sous-matrices de contacts, permet ensuite de déterminer les points de coupure optimaux (Matthews 1976). Le découpage se poursuit de façon récursive jusqu'à ce que la compacité des UPs atteigne une limite minimale fixée par l'utilisateur. Ainsi, la compacité des UPs est quantifiée par un coefficient *R* basé sur le calcul de l'entropie mutuelle des sous-matrices de contacts (Hazout 2007). *R* mesure l'indépendance entre les UPs. Il est maximal (égale à 1) si aucun contact n'existe entre les UPs.

Notre étude sur les contacts au sein des protéines exposée dans le paragraphe précédent, nous a permis de mettre en évidence des spécificités en fonction des atomes pris en compte pour définir un contact mais également des spécificités en fonction de la taille de la protéine ainsi que des biais dans les méthodes de positionnement des chaînes latérales. Dans une seconde étude, à nouveau effectuée avec Guilhem Faure lors de son stage de master 1, nous avons étendu notre analyse à l'organisation hiérarchique des protéines. Ainsi, nous nous sommes intéressés aux contacts observés entre et au sein des UPs définies par le *Protein Peeling* (Faure et al. 2009) (Article 4). Les résultats ont été comparés aux contacts observés dans les protéines entières.

La distribution des acides aminés et des structures secondaires dans les UPs présente peu de différence avec la distribution de référence dans la banque de données de structures entières. De façon étonnante, de même, les préférences de contacts observées entre acides aminés au sein des UPs, ne divergent pas de celles observées dans la banque de référence. Ainsi, du point de vue des contacts, notre étude montre que le *Protein Peeling* définit de véritables unités protéiques, intermédiaires entre les structures secondaires et les domaines et, présentant des caractéristiques similaires à ces derniers. Une hypothèse intéressante est que les contacts observés au sein UPs définit par le *Protein Peeling* se forment durant les premières étapes du processus de repliement, avant la formation de la structure tertiaire du cœur des protéines. Une comparaison avec des résultats expérimentaux permettrait de valider ou d'infirmer cette hypothèse. Cependant, peu de données sont actuellement disponibles pour la communauté

scientifique. Li et collaborateur ont récemment développé une méthode de prédiction spécialement dédiée à l'identification des noyaux de repliement à partir des réseaux de contacts observés au sein des protéines (Li et al. 2008). Leurs résultats ont été évalués sur six protéines dont le mécanisme de repliement est connu et semblent prometteurs. Une comparaison du découpage hiérarchique des protéines fourni par le *Protein Peeling* avec *(i)* les résultats expérimentaux pour ces six protéines d'une part et d'autre part avec *(ii)* les résultats de leur méthode de prédiction pourrait constituer une première étude intéressante.

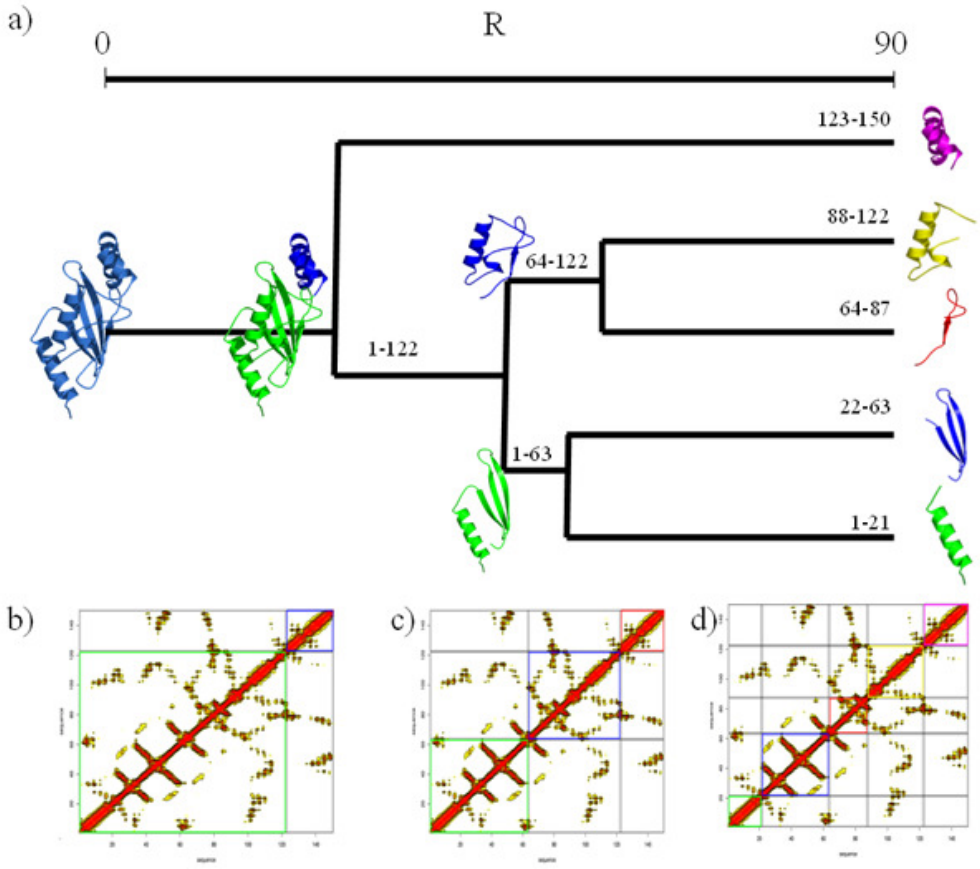

Figure 19. Exemple du déroulement de l'algorithme du Protein Peeling pour l'enzyme de conjugaison à l'ubiquitine d'Arabidopsis Thaliana (code PDB 2aak).

a – La protéine est tout d'abord découpée en deux Unités Protéiques. La première comprend les résidus de 1 à 122, la seconde correspond à la région C-terminale allant du résidu 123 au résidu 150. Dans la seconde étape du processus de découpage, la partie N-terminale est à nouveau découpée en deux UPs de tailles équivalentes (1-63 et 64-122). L'UP 1-63 est ensuite redécoupée en deux UPs : l'UP 1-21 est composée d'une hélice ayant peu de contacts avec la seconde UP 22-63 constituée d'un feuillet β à 3 brins. Enfin, l'UP 64-122 est également coupée en deux UPs : 64-87 et 88-122.
b à d – Matrices de contacts utilisées lors du déroulement de l'algorithme, b – après le premier découpage, c – après le second et d – avec la délimitation des matrices de contacts intra et inter UPs. Figure extraite de (Faure et al. 2009).

2.3.7 Influence de la structure tridimensionnelle sur la conformation locale du squelette polypeptidique (Article 5)

Ainsi, les contacts entre résidus sont le ciment permettant la formation et la stabilisation de la structure 3D des protéines. Par ailleurs, les interactions entre résidus à longues distances appliquent également des contraintes au niveau du squelette polypeptidique et peuvent ainsi influencer la conformation locale. Nous avons vu dans le paragraphe 2.2.6 que du fait de leur structure chimique, les acides aminés ont des préférences marquées pour certaines structures secondaires. Cette observation est la base des méthodes de prédiction des structures secondaires à partir de la séquence. Néanmoins, des études expérimentales ont montré l'importance des interactions à non-locales pour guider la formation d'hélice α ou de brin β (Minor and Kim 1996).

En 1984, Kabsch et Sander ont identifié au sein des protéines des fragments de séquence identiques de longueur limitée pouvant être observés à la fois dans des hélices ou dans des brins (Kabsch and Sander 1984) (voir Figure 20). Ces séquences sont dites *caméléons*. Depuis, de nombreux nouveaux exemples ont été observés et ont confirmé cette première étude (Guo et al. 2007). Ces séquences seraient riches en résidus non-polaires aliphatiques. Elles n'auraient pas de préférence propre pour une conformation en hélice ou en brin (Mezei 1998). Leur conformation est donc largement influencée par des interactions non-locales. Néanmoins, l'information contenue dans les régions flanquantes de la séquence caméléon semble déterminante et permet une prédiction des structures secondaires avec une bonne précision (Jacoboni et al. 2000; Guo et al. 2007).

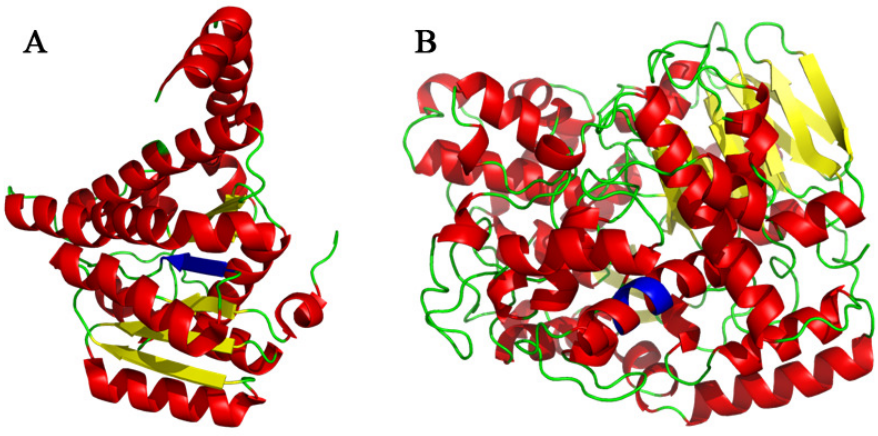

Figure 20. Exemple de séquence caméléon.

Le fragment de séquence MLIL est observé : A – au sein d'un brin β de la 11 beta-hydroxystéroïde deshydrogénase de type I de cochon d'inde (code PDB 1xse) et B – dans une hélice α dans l'aldéhyde ferrodoxine oxidoréductase de *Pyrococcus furiosus* (code PDB 1aor). Figure extraite de (Ghozlane et al. 2009).

Lors du stage de master 1 d'Amine Ghozlane, nous avons récemment étendu les analyses disponibles en nous intéressant à la propension de ces séquences caméléons à prendre des conformations non-régulières de boucles (Ghozlane et al. 2009) (Article 5). Dans une base de données redondante de 40 000 structures protéiques environ (14 692 070 acides aminés), nous avons identifié respectivement 63 228, 34 408, 2 423, 179 et 64 séquences constituées de 4, 5, 6, 7 et 8 résidus. Nous avons par ailleurs estimé les préférences en structures secondaires des séquences de longueur 4. Nos résultats montrent qu'il n'existe pas de préférence marquée pour les conformations α ou β. Néanmoins, pour 22 % des séquences, l'état boucle est défavorisé alors que pour 28 % d'entre elles, celui-ci semble favorisé. Ces spécificités restent cependant peu marquées et sont avant tout dues à la diversité des séquences caméléons. Notre étude est donc en faveur d'une influence importante des facteurs non-locaux pour la détermination de la conformation de ces séquences.

2.3.8 Prédiction de la structure tridimensionnelle à partir de la séquence protéique

Lorsque la structure tridimensionnelle d'une protéine n'est pas connue, la construction d'un modèle 3D à partir de sa séquence en acides aminés peut fournir des informations essentielles sur ses propriétés et son mécanisme de fonctionnement. Néanmoins, les protéines sont des molécules « géantes ». Le nombre considérable de degrés de liberté de la chaîne polypeptidique et le nombre non moins fabuleux d'interactions possibles entre ses atomes font de la construction de modèles protéiques un défi sensationnel encore non résolu.

Dès 1966, alors que seulement trois structures cristallographiques étaient résolues (pour la myoglobine, l'hémoglobine et le lysozyme), Levinthal proposa un modèle pour un segment de 14 résidus appartenant au cytochrome C (Levinthal 1966) (voir Figure 21). En se basant sur la puissance de calcul d'ordinateurs récents conçus au Massachussetts Institute (MIT), il propose un logiciel permettant de construire des modèles protéiques en se basant à la fois sur l'expertise du biochimiste et sur les capacités de calcul des machines. La chaîne polypeptidique est plongée dans un champ de force permettant de calculer l'énergie du système. Ainsi, le chercheur a la possibilité d'appliquer les déformations lui paraissant pertinentes pour créer un modèle 3D. Une visualisation tridimensionnelle de la molécule est projetée sur un écran d'oscilloscope. Les rotations sont commandées grâce à une boule de commande (*trackball* en anglais) et les déformations grâce à un stylet lumineux (cf. Figure 21). Du fait, des capacités de calcul limitées des ordinateurs à cette époque, ce travail n'a pas

eu un impact immédiat. Néanmoins, grâce à ses idées incroyablement novatrices, Levinthal est le précurseur des travaux réalisés aujourd'hui dans le domaine de la modélisation *in silico* des structures protéiques.

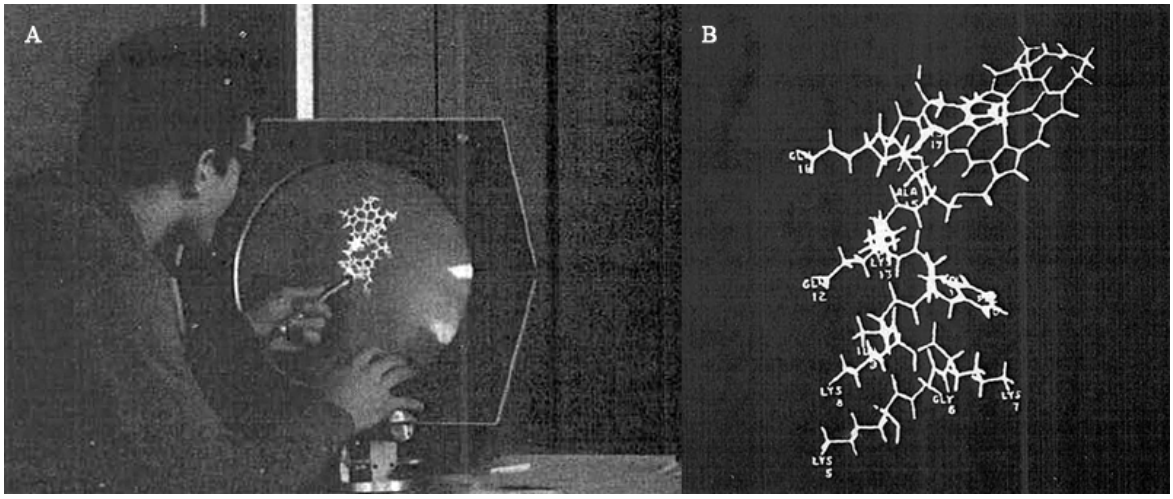

Figure 21. Construction d'un modèle pour un segment du Cytochrome C par Levinthal en 1966.

A – Unité de visualisation des molécules développée par John Ward et Robert Stotz et couplés aux unités centrales du MIT. La communication avec l'ordinateur se fait *via* le clavier et le stylet lumineux (main gauche). La rotation de la molécule est dirigée par la boule de commande (main droite). B – Modèle moléculaire du segment du Cytochrome C allant du résidu 5 au résidu 18. Le modèle soutient l'hypothèse d'une conformation en hélice α. Figure adaptée de (Levinthal 1966).

Aujourd'hui, les méthodes de prédiction de la structure tridimensionnelle des protéines peuvent être classées en trois catégories :

- La modélisation par homologie (ou comparative).
- Les méthodes de reconnaissance de repliement (ou d'enfilage, *threading* en anglais).
- Les méthodes *ab initio* et *de novo*.

Le choix de la méthode est fonction des informations disponibles pour la réalisation du modèle. Le facteur le plus déterminant dépend de l'existence ou non dans la PDB d'une structure protéique résolue de séquence similaire à celle de la protéine à modéliser et du taux d'identité de séquence entre ces protéines. L'idée sous-jacente est l'identification de protéines homologues ayant conservé une structure proche de la séquence cible malgré une divergence des séquences due au processus d'évolution moléculaire. La Figure 22 résume les taux d'identité de séquence requis entre la séquence cible et la (ou les) protéine(s) servant de modèle(s) pour chacune des catégories des méthodes de modélisation. Le niveau de résolution obtenu ainsi que les applications possibles des modèles sont également indiqués.

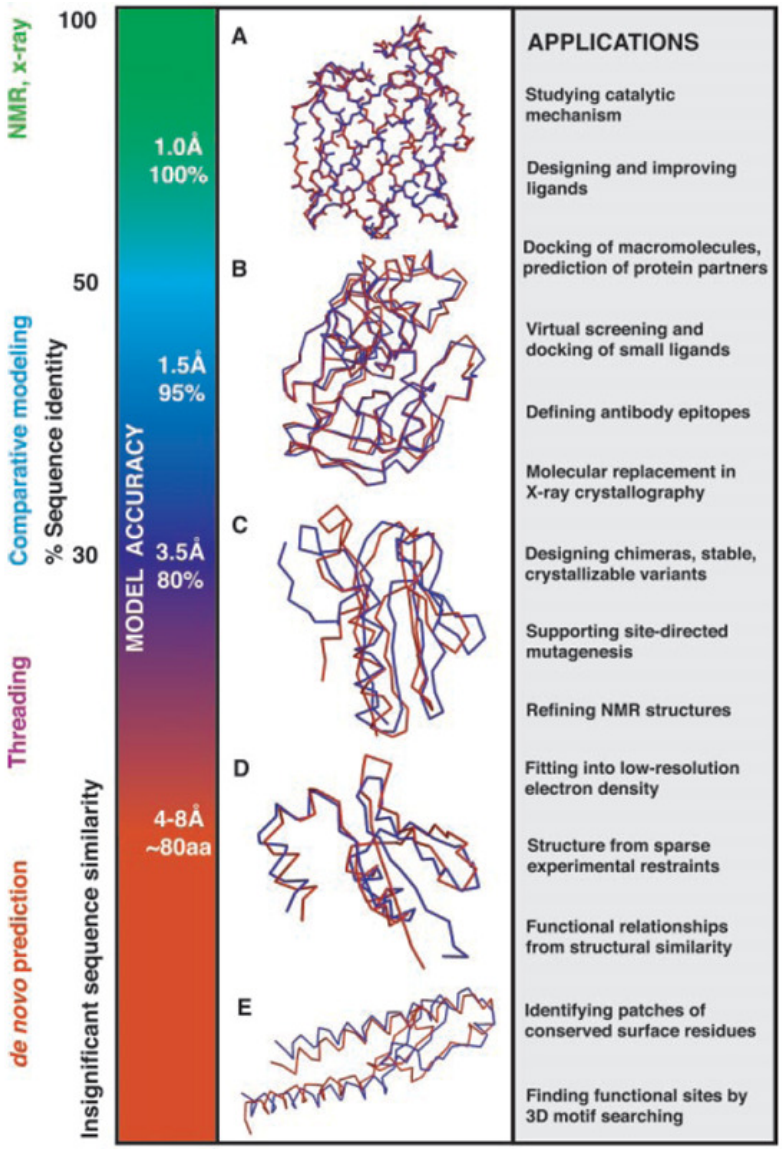

Figure 22. Méthodes de détermination de la structure 3D des protéines : identités de séquence requises avec les structures supports, niveaux de résolution et applications.

Figure extraite de (Baker and Sali 2001).

2.3.8.1 Modélisation par homologie

La modélisation par homologie requiert l'existence d'une protéine de structure connue présentant plus de 30% d'identité de séquence avec la protéine à modéliser (Baker and Sali 2001). Cette structure servira de structure de référence.

Le logiciel le plus largement utilisé de nos jours est Modeller crée par Sali et Blundell en 1993 (Sali and Blundell 1993). La modélisation se fait en plusieurs étapes : (1) la sélection de la structure de référence la plus pertinente biologiquement, (2) l'alignement des séquences de la protéine cible et du support, (3) la construction du modèle à proprement parlé en s'appuyant sur les régions conservées. Dans un second temps, les régions moins conservées, notamment les boucles, sont modélisées et les chaînes latérales positionnées (voir paragraphe

2.3.5). Le modèle doit ensuite être raffiné et validé en fonction des connaissances et notamment de données expérimentales non utilisées pour sa construction.

Actuellement, cette méthode est celle permettant d'obtenir les modèles les plus précis. Cependant, comme nous l'avons vu (paragraphe 2.3.5), le positionnement des chaînes latérales est encore délicat. De même, la modélisation des boucles, plus flexibles et de séquence souvent plus divergente par rapport à la structure de référence, reste un problème très difficile. Les algorithmes de prédiction les plus récents se concentrent sur la résolution concomitante de deux tâches : *(i)* l'échantillonnage de l'espace conformationnel du fragment de structure et *(ii)* l'évaluation des conformations échantillonnées (Prédiction des boucles dans Modeller (Fiser et al. 2000), LOOPY (Xiang et al. 2002), RAPPER (de Bakker et al. 2003), PLOP (Zhu et al. 2006), *LoopBuilder* (Soto et al. 2008)). L'échantillonnage peut être réalisé à partir de banques de boucles ou *ab initio*. Cette dernière solution semble être la plus performante pour les boucles longues (Xiang et al. 2002). La plupart des méthodes récentes sont relativement efficaces pour les boucles de moins de 10 résidus. Au dessus de cette limite, l'étape limitante est l'échantillonnage car celui-ci requiert des temps de calcul considérables pour être efficace et permettre une modélisation précise (Zhu et al. 2006). La modélisation des boucles est un champ de recherche encore très actif.

2.3.8.2 Méthodes de reconnaissance de repliement (« Threading »)

Lorsque le taux d'identité avec les séquences disponibles dans la PDB descend au dessous de 30% mais reste supérieur à 15%, les méthodes de reconnaissance de repliement sont une alternative à la modélisation par homologie. Le principe de ces méthodes est d'identifier parmi les repliements connus celui que pourrait adopter la séquence cible. De manière imagée, la séquence à modéliser est « enfilée » dans des repliements connus. Un score de compatibilité est ensuite calculé afin de déterminer la meilleure hypothèse.

Les principales difficultés de cette approche résident dans l'alignement séquence-structure et dans l'évaluation des différents modèles potentiels. Elle est de plus limitée par le nombre de repliements connus. Dernièrement, ces méthodes sont devenues de moins en moins compétitives avec les méthodes dites *ab initio* et surtout *de novo* (Moult 2005).

2.3.8.3 Méthodes *ab initio* et *de novo*

Les méthodes *ab initio* visent à prédire la structure d'une protéine à partir de la seule connaissance de sa séquence en acides aminés. Cet objectif est un défi scientifique majeur. Cependant, dans le contexte des projets de séquençage à grande échelle générant un grand

nombre de séquences sans aucune structure homologue connue, le développement de ces approches est particulièrement pertinent.

Deux types de méthodes sont à distinguer : les méthodes *ab initio* « pures » et les méthodes *de novo*. Les méthodes *ab initio* « pures » reposent uniquement sur les propriétés physico-chimiques des protéines et sur la recherche de la structure de plus basse énergie parmi l'ensemble des structures possibles. Les méthodes les plus récentes utilisent une approche hiérarchique comme dans ASTRO-FOLD dans laquelle les hélices sont prédites en premier, puis les brins et la topologie globale des feuillets. Les boucles sont modélisées à part. L'ensemble des contraintes issues de ces prédictions isolées sont enfin utilisées pour la prédiction d'une structure 3D (Klepeis and Floudas 2003). Ces méthodes nécessitent une exploration de l'espace conformationnel très importante. Elles restent donc pour le moment applicables uniquement à de petites protéines de moins d'une centaine de résidus.

Les méthodes dites *de novo* exploitent les informations obtenues de l'analyse des structures tridimensionnelles connues. Actuellement, les approches les plus performantes sont les méthodes d'assemblage de fragments (Moult 2005). Ces méthodes reposent sur l'hypothèse selon laquelle, même si nous n'avons pas encore observé tous les repliements possibles, nous avons probablement vu presque toutes les sous-structures. Étant donné que la relation séquence-structure n'est pas suffisamment forte pour déterminer avec certitude la structure de fragments de séquence, la première étape consiste à réduire le nombre de conformations possibles pour un segment de séquence donné. Ces conformations sont extraites d'une banque des structures locales connues. Les fragments prédits sont ensuite assemblés pour générer de nombreux modèles 3D possibles. L'exploration de l'espace conformationnel est beaucoup plus rapide que pour des méthodes *ab initio* pures grâce aux contraintes locales imposées par les fragments prédits. Le défi réside ensuite dans la sélection du modèle le plus pertinent grâce à une fonction d'énergie appropriée. La méthode remportant le plus de succès depuis plusieurs années est ROSETTA développée par le groupe de Baker (Rohl et al. 2004; Jauch et al. 2007). Les auteurs considèrent des fragments de 3 à 9 résidus de long et réalisent une exploration conformationnelle très large grâce à de très fortes capacités de calculs.

2.3.8.4 Evaluation des méthodes et des modèles

L'évaluation et la comparaison des différentes méthodes de prédiction est une tâche ardue. Ainsi, tous les deux ans, la communauté se réunie pour une compétition nommée CASP (*Critical Assessment of Structure Prediction*). Les méthodes participantes sont comparées sur un même jeu de protéines et avec des critères identiques. Les structures protéiques à prédire

ont été déterminées expérimentalement très récemment et ne sont pas encore déposées dans la PDB ou publiées. La dernière édition, CASP8, a eu lieu en 2008 et a évalué les performances de 233 méthodes de prédiction entièrement automatisées ou non, sur 128 protéines cibles.
L'évaluation des performances des méthodes est réalisée en comparant les modèles aux structures 3D obtenues expérimentalement. Le choix d'un critère mesurant la similitude ou distance entre ces deux structures n'est pas trivial. Classiquement, le critère utilisé est l'écart quadratique moyen (*Root Mean Square Deviation* ou RMSD en anglais). Il correspond à la distance euclidienne entre les coordonnées des atomes des deux structures après superposition optimale :

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i^s - r_i^t)^2}$$

Avec *s* et *t* les deux structures comparées, N le nombre total d'atomes comparés, r_i^s et r_i^t les coordonnées de l'atome *i* des structures *s* et *t* respectivement. Le RMSD est souvent calculé uniquement sur les atomes du squelette polypeptidique voire seulement sur les Cα. La limitation principale de cette mesure est sa dépendance vis-à-vis de l'alignement initial des structures à comparer. En effet, l'alignement des structures 3D protéiques est encore actuellement un champ de recherche à part entière et devient un problème difficile si les structures diffèrent significativement. Ainsi, d'autres critères peuvent être utilisés. A CASP notamment, le GDT_TS (*Global Distance Test Total Score*) est privilégié. Il repose sur différents seuils de distance entre Cα et identifie le plus grand ensemble de résidus pouvant être alignés à la structure de référence en considérant un seuil donné (Zemla et al. 1999).

2.3.9 Apport des structures 3D pour l'annotation fonctionnelle des protéines (Article 6)

La structure 3D des protéines est classiquement considérée comme le support de leur fonction biologique. Ainsi, elles apportent des informations essentielles pour mieux comprendre le mécanisme de fonctionnement des protéines, leur rôle biologique et donc notamment pour aider à concevoir de nouveaux médicaments. Dans ce dernier cas, si la fonction de la protéine cible est inconnue, la recherche de sites de liaison hypothétiques est une première étape importante. Ces différents sites pourront être explorés ou comparés à des sites connus d'autres protéines (Kroemer 2007). Une autre application est l'utilisation des structures 3D protéiques pour le développement d'Anticorps. En effet, les informations de structure peuvent être

précieuses pour prédire le site de liaison d'un anticorps sur une protéine antigène (*épitope*) à l'aide de méthodes bioinformatiques (Moreau et al. 2006; Moreau et al. 2008). Une fois identifiés ces épitopes peuvent être produits sous forme de peptides et utilisés pour le design et la production d'anticorps.

De même, l'utilisation directe de la structure 3D pour attribuer une fonction est un champ de recherche en plein essor. Ce développement est du l'augmentation du nombre de structures de protéines hypothétiques de fonction inconnue rendues disponibles grâce aux projets de génomique structurale. Les premières méthodes développées reposaient directement sur une identification de similarités locale de structure. SuMo et SiteEngine sont basés sur une description chimique de la surface des protéines et des comparaisons de graphes (Jambon et al. 2003; Shulman-Peleg et al. 2004). ProFunc utilise des méthodes basées à la fois sur la séquence et la structure (Laskowski et al. 2005b). Ces méthodes sont complémentaires des approches classiques basées sur la séquence. En effet, ces dernières ne permettent d'annoter qu'un peu plus de la moitié des protéines de fonction inconnue (Doppelt et al. 2007).

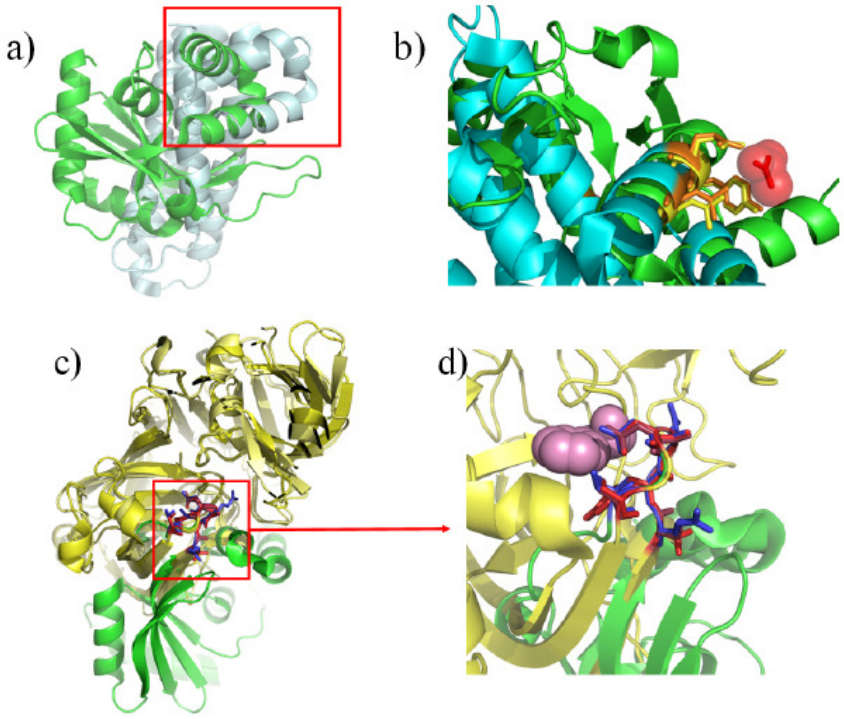

Figure 23. Exemples de résultats obtenus avec MED-SuMo.

Une protéine hypothétique est analysée (vert) (code PDB 2EWR). MED-SuMo identifie deux régions d'intérêt présentées d'une part en a/b et d'autre part en c/d. (a-b) Le site de liaison similaire entre la protéine cible et la protéine 2CJ5, inhibitrice d'invertases, (a) superposition complète des structures autour du site de liaison, (b) zoom. (c-d) Le site de liaison similaire entre la protéine cible et la protéine 5APR, Hydrolase, (c) superposition des deux structures autour du site de liaison, (d) zoom sur la région d'intérêt. Figure tirée de (Doppelt et al. 2007).

Ce logiciel a été amélioré par la société MEDIT-SA (http://www.medit-pharma.com/) sous le nom de MED-SuMo. Plus rapide, il permet des recherches de similarités à plus grande échelle. J'ai ainsi été associée à une étude menée par Olivia Doppelt lors de sa thèse (Doppelt et al. 2007) (Article 6). L'étude avec MED-SuMo, d'une protéine hypothétique résolue par le consortium « *Joint Center for Structural Genomics* » (JCSG, http://www.jcsg.org/) a permis de détecter deux surfaces d'interaction similaires sur des protéines connues alors qu'aucune des méthodes d'annotation classiques n'avait fourni de résultat exploitable (Figure 23).

2.4 Structure quaternaire et assemblage moléculaire

Une protéine fonctionnelle peut-être constituée d'une seule chaîne polypeptidique ou d'un assemblage de plusieurs chaînes (identiques ou non). Cet assemblage est alors nommé *structure quaternaire* et chaque chaîne polypeptidique est un *monomère*. Les protéines ayant une structure quaternaire sont dites *multimèriques*. L'association des monomères est principalement stabilisée par des interactions non-covalentes (paragraphe 2.3.4).

2.5 Conclusion

Les protéines sont des molécules *géantes* constituées d'un enchaînement linéaire d'acides aminés formant une séquence. Cette séquence se replie en une structure 3D fonctionnelle, spécifique et stabilisée par des interactions entre résidus à courte et à longue distance. La structure 3D peut être obtenue expérimentalement. Néanmoins, actuellement, la détermination expérimentale des structures protéiques fait face à de nombreuses difficultés techniques et à des coûts élevés. Aussi, la prédiction *in silico* de la structure tridimensionnelle complète d'une protéine à partir de sa séquence en acides aminés est une alternative et constitue un défi scientifique d'un intérêt majeur. Pour parvenir à cet objectif, les méthodes *de novo* procédant par assemblage de fragments semblent les plus pertinentes. Ainsi, une description efficace des structures protéiques et de leur architecture ainsi qu'une caractérisation précise de la relation structure-séquence sont essentielles.

Classiquement, les structures 3D sont analysées en termes de structures secondaires. Depuis ces 60 dernières années, leur caractérisation a fait l'objet de très nombreuses études et cette description a également été largement utilisée par la communauté scientifique. Leur géométrie est dépendante des degrés de liberté autorisés par les propriétés du squelette polypeptidique.

De plus, elles sont fréquemment stabilisées par des interactions locales, à l'exception des feuillets β stabilisés par des interactions à plus longue distance. De fortes spécificités de séquence existent donc et permettent le plus souvent une prédiction très satisfaisante. Cependant, il faut de noter que la « simple » assignation des structures secondaires à partir de la structure 3D n'est pas triviale. De nombreux désaccords existent entre les méthodes. Il est important d'en avoir conscience pour mieux comprendre des divergences éventuelles entre analyses. De plus, le plus souvent, seuls trois états sont considérés (Hélice α, Brin β et boucles). Cette description n'est pas suffisante pour caractériser les structures tridimensionnelles dans leur ensemble. L'état boucle notamment exclut les états Hélice et Brin mais n'est pas réellement informatif quant à la conformation de squelette polypeptidique, alors qu'il représente près de 50% des résidus. De même, comme nous l'avons vu, les états Hélices et Brins rassemblent également des conformations assez hétérogènes. Par ailleurs, les structures secondaires décrivent les structures résidu par résidu, elles ne fournissent aucune information concernant l'orientation des éléments les uns par rapport aux autres.
Une description plus fine des structures est donc nécessaire.

3. CARACTÉRISATION FINE ET PRÉDICTION DES STRUCTURES LOCALES PROTÉIQUES

L'observation selon laquelle les protéines sont constituées d'un répertoire limité de structures locales récurrentes (Fitzkee et al. 2005; Banavar and Maritan 2007) a conduit au développement de librairies de fragments ou *alphabets structuraux* (Offmann et al. 2007). Ces alphabets structuraux permettent de dépasser les limites des structures secondaires (paragraphe 2.5). Ils proposent une caractérisation unifiée de la structure de fragments protéiques et décrivent la conformation de tous les résidus. Le but de leurs concepteurs est d'identifier de la manière la plus optimale possible ces structures locales récurrentes. Dans ce contexte, les protéines sont vues comme des assemblages de *blocs structuraux élémentaires*. Ainsi, la prédiction de la structure 3D globale des protéines peut être décomposée en un problème hiérarchique. Dans un premier temps, il est nécessaire de parvenir à prédire la conformation adoptée par des fragments de séquence. L'assemblage des fragments de structure est le défi suivant à relever.

Dans ce chapitre, je présenterai tout d'abord une synthèse des nombreuses études réalisées sur ce thème. Je consacrerai ensuite deux paragraphes (3.2 et 3.3) à la description des approches développées au sein du laboratoire : l'alphabet composé par les Blocs Protéiques et son extension, la Bibliothèque des *Prototypes de Structures Locales* (PSLs). Cette présentation est importante pour une bonne compréhension des chapitres suivants présentant mes principaux travaux de thèse. En effet, ces derniers s'articulent autour de deux thématiques principales : l'optimisation de la prédiction des PSLs (section 4) et l'étude de la flexibilité des PSLs au sein des structures protéiques (section 6).

3.1 Notion de bibliothèques de fragments

3.1.1 Caractéristiques et utilisations des bibliothèques de fragments

De nombreuses bibliothèques de fragments structuraux ont été développées et ont donné lieu à de multiples applications. Elles présentent de nombreuses différences comme par exemple le nombre de groupes structuraux, la longueur des fragments considérés ou encore la méthode de conception utilisée. Toutefois, la démarche globale reste identique. Le Tableau 4 présente un résumé des bibliothèques existantes.

Tableau 4. Synopsis des différentes librairies de structures locales et des alphabets structuraux.

Equipe	Année	*Nom de la librairie*	Nombre de protéines utilisées	Nombre de fragments utilisés	Méthode d'apprentissage	Descripteurs et Distances utilisés	Nombre de prototypes	Longueur des prototypes
Unger *et al.*	1989	*Building Blocks*	4\82	426\12 973	*k* -moyennes	Cα RMSD	103	6
Rooman *et al.*	1990	*Recurrent local structural motifs*	75	12 978	Classification hiérarchique	Cα RMSD	4	4, 5, 6 et 7
Prestrelski *et al.*	1992	*Substructures*	14	2 347	Fonction	Distance linéaire et angle α	113	8
Zhang *et al.*	1993	*Structural Building Blocks*	74	13 114	AutoANN et *k* -moyennes	Distances entre Cα, angles diédres et de valence	6	7
Schuchhardt *et al.*	1996	*Local structural motifs*	136	24 239	Carte topologique de Kohonen	Angles dièdres	100	9
Fetrow *et al.*	1997	*Structural Building Blocks*	116	23 335	AutoANN et *k* -moyennes	Distances entre Cα, angles diédres et de valence	6	7
Bystroff et Baker	1998	*I-sites*	471	NA	*k* -moyennes	Profiles de séquence et RMSD / dma	13 à partir de 82 (réévaluées à 16 en 2000)	Structure : 3 à 15, Séquence : 8
Camproux *et al.*	1999	*Short Structural Building Blocks*	100	19 137	HMM	Distances entre Cα	12	4
Micheletti *et al.*	2000	*Oligons*	75	11 086	Classification itérative en supprimant progressivement les clusters les plus importants	Cα RMSD	28, 202, 932 et 2 561	4, 5, 6 et 7
de Brevern *et al.*	2000	*Protein Blocks*	342	87 996	Classification non supervisée (~SOM + transitions)	Angles dièdres	16	5
Kolodony *et al.*	2002	-	145\200	NA (~5 000 to 9 000)	*k* -moyennes + Recuit simulé	Cα RMSD	4 à 14 (20 dans Le *et al.* , 2009), 10 à 225,40 à 300, 50 à 250	4, 5, 6 et 7
Hunter et Subramaniam	2003	*Centroids*	790	156 643	*k* -moyennes + Recuit simulé	Différence entre fragments dans un espace hypercosine	28 à 16336 (28 pour la prédiction)	7
Camproux *et al.*	2004	*Short Structural Building Blocks*	250 x 2	*NA*	HMM	Distances entre Cα	27	4
De Brevern, Etchebest *et al.*	2005	*Protein Blocks*	1 407	293 507	Nouvelle évaluation	Angles dièdres	16	5
Benros *et al.*	2006	*LSP*	675 et 1 401	139 503 et 251 497	Modèle de la Protéine hybride	Blocs Protéiques et Ca RMSD	120	11
Sander *et al.*	2006	*Structural representatives*	1 999	295 411	Algorithme du Leader et k-moyennes	Matrice de distance entre Cα	28	7
Tung *et al.*	2007	*Kappa-alpha*	1 348	225 523	Classification selon la méthode du plus proche voisin	Angles κ et α, Cα RMSD	23	5
Dong *et al.*	2007	-	1 400	179 463 pour l'apprentissage	*k* -moyennes	Cα RMSD	28	7
Yang	2008	*Protein Folding Shape Code (PFSC)*	200 pour l'évaluation	NA	Détermination de 3 seuils sur les 3 descripteurs sans partition des zones associées à un type de structure secondaire (DSSP)	Angles de torsion entre les 4 premiers et les 4 derniers Cα, Distance entre les Cα des extrémités	27	5
Ku *et al.*	2008	*Centroids*	NA	20 953 584	Carte topologique de Kohonen et *k* -moyennes	Angles dièdres	18	5
Dong *et al.*	2008	*Building Blocks*	1 219	NA	Méthode du Protein Peeling *k* -moyennes	Angles de torsion, profils de séquences	43, 50, 51 et 36 => 180 au total	4, 5, 6 et 7

Les librairies indiquées avec un fond coloré sont associées à une méthode de prédiction des structures locales. Tableau adapté de (Offmann et al. 2007) et mis à jour.

Les bibliothèques de fragments sont généralement construites selon la démarche suivante : *(i)* les auteurs créent tout d'abord une banque de structures cristallographiques non redondantes représentative des structures protéiques connues, *(ii)* les structures sont ensuite découpées en

fragments de longueur *N*, puis finalement, *(iii)* les fragments sont regroupés en fonction de leurs similitudes en *X* groupes et un fragment représentant chaque groupe est désigné.

Les différences principales entre bibliothèques de fragments s'expliquent par des différences d'objectifs. Certaines bibliothèques ont pour objectif de décrire de manière précise les structures protéiques. Elles permettent de reconstruire les structures par assemblage de fragments (Micheletti et al. 2000; Kolodny et al. 2002 ; Camproux et al. 2004; Dong et al. 2007), de les caractériser pour les classer (Schuchhardt et al. 1996; Le et al. 2009) ou encore les comparer (Tung et al. 2007; Ku and Hu 2008). Ces dernières sont généralement constituées d'un nombre important de groupes structuraux et considèrent souvent des fragments relativement longs. Les travaux pionniers de Unger et collaborateurs ont par exemple abouti à une librairie de 103 classes structurales pour des fragments de 6 résidus (Unger et al. 1989). Schuchhardt *et al.* ont également développé une bibliothèque de 100 classes de fragments structuraux de 9 résidus de long (Schuchhardt et al. 1996).
D'autres bibliothèques sont dédiées à la prédiction des structures locales à partir de la séquence en acides aminés (voir les lignes orangées du Tableau 4). Un difficile équilibre doit alors être trouvé entre une bonne description des structures et la mise en évidence de relations séquence-structure pertinentes pour la prédiction. Dans ce contexte, le nombre de classes structurales est généralement moins important et les fragments considérés moins longs. de Brevern *et al.* développent un alphabet de 16 classes de fragments de longueur 5 (de Brevern et al. 2000). De même, Hunter et Subramaniam utilisent 28 classes de fragments de 7 résidus (Hunter and Subramaniam 2003). Il convient dans ce cadre de noter le nombre important de classes et la longueur des fragments de la bibliothèque développée par Benros *et al.* : des fragments de 11 résidus et 120 classes sont considérés.

Les différences entre bibliothèques sont également dues à des choix propres aux auteurs.
Ainsi, la longueur des fragments peut être fixe (Unger et al. 1989; de Brevern et al. 2000; Camproux et al. 2004; Benros et al. 2006; Ku and Hu 2008) ou variable (Bystroff and Baker 1998; Dong et al. 2008), l'analyse étant plus difficile dans le second cas. De même, les fragments sont le plus souvent chevauchants mais ils peuvent également ne pas l'être (Kolodny et al. 2002; Dong et al. 2008). Considérer des fragments chevauchants présente l'avantage de permettre la prise en compte des relations de dépendance existantes entre fragments successifs le long des protéines. Toutefois, pour certains auteurs, ce chevauchement rend la classification plus difficile (Dong et al. 2008).

Par ailleurs, une grande variété existe au niveau de la caractérisation des fragments. Ils peuvent être décrits en coordonnées cartésiennes (les coordonnées des Cα sont alors le plus souvent considérées) (Unger et al. 1989; Micheletti et al. 2000) ou en coordonnées internes (angles Φ/Ψ, α[3], κ[4] ;(de Brevern et al. 2000; Tung et al. 2007)). Ils sont alors le plus souvent comparés en terme de Cα RMSD (paragraphe 2.3.8.4) ou RMSD angulaire (paragraphe 3.2.1). Des distances entre Cα peuvent également être considérées (Camproux et al. 1999; Sander et al. 2006; Yang 2008). Par exemple, Camproux *et al.* décrivent un fragment grâce à quatre distances (Figure 24A). Les trois premières distances correspondent aux trois distances entre Cα non consécutifs, *i.e.,* entre le premier et le 3ème Cα (d_1), entre le 1er et 4ème (d_2), et entre le 2ème et 4ème (d_3). La dernière distance est la projection du 4e Cα dans le plan des trois premiers. Fetrow et collaborateurs utilisent une méthode plus complexe : ils encodent tout d'abord les fragments grâce à 43 descripteurs de distances et d'angles puis résument cette information dans des vecteurs de longueur 8 grâce à un réseau de neurones auto-corrélé, finalement, ces vecteurs sont l'information utilisée pour classer les fragments structuraux (voir Figure 24B) (Fetrow et al. 1997). Nous verrons également, dans le paragraphe 3.3.1, que Benros *et al.* utilisent une méthodologie tout à fait originale : les fragments sont décrits par des séries de Blocs Protéiques avant d'être classés (Benros et al. 2006). Enfin, la séquence en acides aminés est aussi parfois prise en compte dans la description des fragments (Bystroff and Baker 1998; Bystroff et al. 2000; Dong et al. 2008). Cependant, dans ce cas, l'équilibre entre une faible variabilité structurale *intra*-classe et une forte similarité de séquence peut être difficile à trouver.

Les méthodes utilisées pour classer les fragments sont également très diverses. Les k-moyennes (*k-means* en anglais) sont très souvent utilisées (Unger et al. 1989; Fetrow et al. 1997; Bystroff and Baker 1998; Kolodny et al. 2002; Dong et al. 2007; Dong et al. 2008). Mais, il convient également de citer la classification hiérarchique (Rooman et al. 1990), les cartes topologiques de Kohonen (Schuchhardt et al. 1996; de Brevern et al. 2000) ou encore les chaînes de Markov Cachées (Camproux et al. 2004). Par exemple, après avoir trié les fragments de 5 résidus de long selon leurs angles α et κ, Tung et collaborateurs, utilisent la méthode du plus proche voisin en se basant sur le critère du Cα RMSD (voir la Figure 25) (Tung et al. 2007).

[3] L'angle α d'un résidu *i* est l'angle dièdre formé par les quatre Cα *i-1, i, i+1 et i+2*. Il varie entre -180° à +180°.

[4] L'angle κ d'un résidu *i* est formé par les trois Cα des résidus *i-2, i et i+2*.

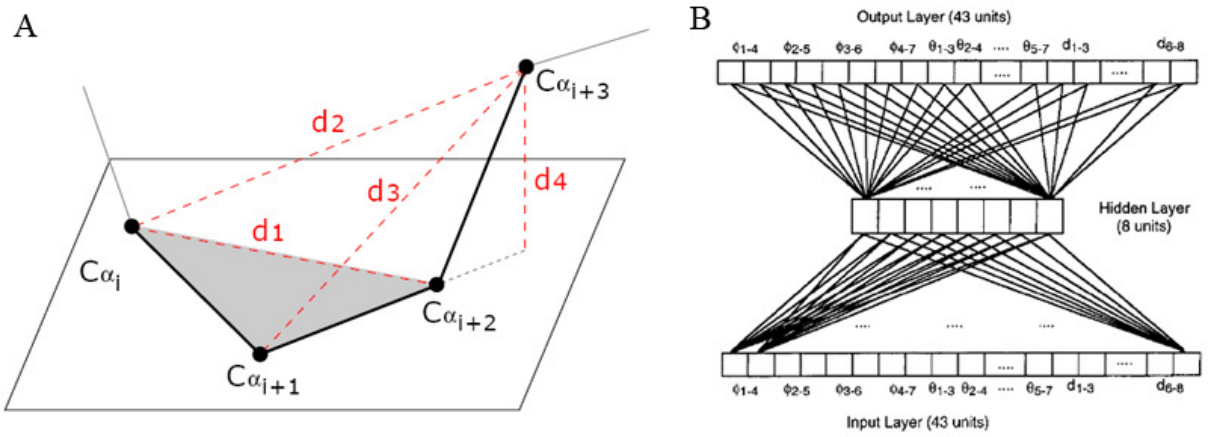

Figure 24. Exemple de description des fragments de structures locales.

A – Descripteurs utilisés par Camproux et collaborateurs (Camproux et al. 1999; Camproux et al. 2004). Un fragment de quatre résidus est décrit par les trois distances (d_1, d_2, d_3) entre les carbones α non successifs et la projection d_4 du quatrième carbone α sur le plan formé par les trois autres. B – Description utilisée par Fetrow *et al.* Un réseau de neurone est entrainé pour reproduire une couche de sortie identique à la couche d'entrée de 43 neurones en passant par une couche cachée de 8 neurones. Les valeurs prises par la couche cachée pour chaque fragment sont utilisées pour le décrire. Figure extraite de (Fetrow et al. 1997).

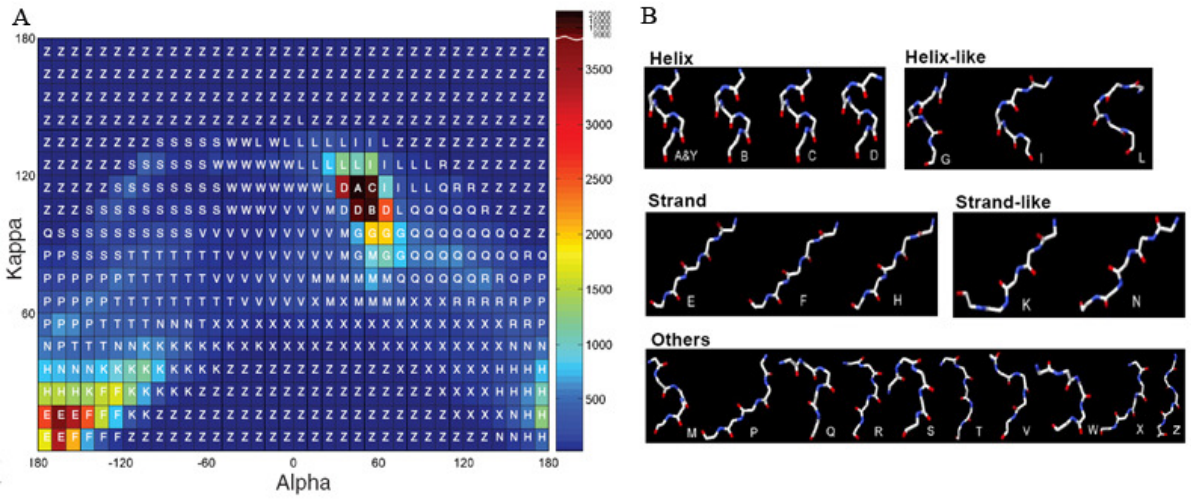

Figure 25. Un exemple de stratégie de classification des fragments avec l'alphabet Kappa-Alpha.

A – Les auteurs caractérisent tout d'abord les fragments de 5 résidus de long grâce à leurs angles α et κ. Les fragments sont ensuite répartis sur la grille présentée ci-dessus discrétisant les espaces des angles en créant une division tous les 10°. Un fragment représentatif est alors désigné pour chaque cellule de la grille. Finalement, les cellules sont regroupées dans une même classe en fonction de la similarité géométrique de leur représentant (Cα RMSD) selon la méthode du plus proche voisin. La cellule la plus peuplée est traitée en premier et est regroupée avec la cellule la plus similaire. Cette étape est répétée tant que la similarité des cellules et le nombre de fragments dans une classe sont inférieurs à des seuils donnés. Les mêmes étapes sont répétées pour les cellules restantes. Le gradient de couleur de la grille correspond au nombre de fragments dans chaque cellule. B – Aperçu des 23 classes obtenues. Les lettres sous les fragments représentatifs des classes correspondent aux lettres sur la grille. Figure extraite de (Tung et al. 2007).

La qualité d'une bibliothèque de fragment doit être évaluée sur les structures protéiques d'un échantillon de validation, indépendant de l'échantillon d'apprentissage. Différents critères sont utilisés. La qualité de l'approximation locale est essentielle. Dans le cas de bibliothèques dédiées à la reconstruction de structures 3D par assemblage de fragments, la qualité de l'approximation globale peut également être évaluée. Ce critère consiste à comparer la structure reconstruite par rapport à la structure 3D réelle avec un calcul de Cα RMSD. Lorsque l'objectif est la prédiction des structures locales, l'analyse de la relation structure-séquence passe généralement par des matrices d'occurrences. Par ailleurs, il est important de noter que certaines bibliothèques de fragments sont nommés « alphabets structuraux » (Bystroff et al. 2000; de Brevern et al. 2000; Camproux et al. 2004). Leurs représentants moyens sont vus comme les lettres d'un alphabet, et la caractéristique mise en avant est la capacité qu'ont ces lettres à s'associer suivant des règles logiques pour former des mots plus longs. La succession des lettres n'est donc pas aléatoire, et toutes les successions ne sont pas possibles (Benros 2005).

Une comparaison des différentes bibliothèques est très complexe du fait de leur faible mise à disposition vis-à-vis de la communauté scientifique, de la diversité des stratégies utilisées et des différentes longueurs des fragments. Le Tableau 5 illustre une comparaison entre les SBBs définis par Fetrow *et al.* et les Blocs Protéiques de de Brevern et collaborateurs (Fetrow et al. 1997; de Brevern et al. 2000). La confusion entre les classes rend la comparaison difficile.

Tableau 5. Comparaison entre les Blocs Protéiques (*Protein Blocks*) de de Brevern et al. et les Blocs Structuraux (*Structural Building Blocks*) de Fetrow et al.

		Protein Blocks															
		a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*
SBBs	η		5.7						**11.4**	1.6	2.2	**11.9**	9.2	**23.1**	**15.4**	**11.7**	2.7
	α												**10.6**	**86.1**			
	τ	7.1	6.4	**10.2**	**16.0**		2.2	5.0		**12.7**			5.2	7.3	1.2	8.2	**16.4**
	ζ	2.0		1.5	1.3	9.4	**34.9**	1.9	5.1		2.1	**27.2**	2.7	9.0			1.1
	ι	7.1	**10.6**	**15.8**	**40.4**	6.5	3.7	2.0	1.9	1.5	1.1	1.0	1.0	1.2			4.9
	β	7.2	7.1	**18.3**	**53.2**	2.4	6.1		1.6								1.9

La comparaison a été réalisée sur la seule bibliothèque de protéines codées en *SBBs* disponible. La fréquence f_{ij} indique le pourcentage fragments assignés au SBB_i associés au PB_j. Seules les fréquences supérieures à 1% sont notées, celles en gras sont supérieures à 10 %. Tableau extrait de (Offmann et al. 2007).

Dans le cadre de la prédiction des structures locales à partir de la séquence, la diversité des stratégies mises en place est là encore très importante. La prédiction peut être basée sur le théorème de Bayes permettant de calculer la probabilité de chaque classe structurale en une position (cf paragraphe 3.2.4) (de Brevern et al. 2000; Hunter and Subramaniam 2003). Des méthodes d'apprentissage plus sophistiquées peuvent également être utilisées comme la régression logistique (paragraphe 3.3.4.1) (Benros 2006), les machines à vecteurs support (ou *SVM*, paragraphe 3.1.2.2) (Sander et al. 2006; Dong et al. 2008) ou encore les forêts d'arbres de décision (Sander et al. 2006). Dong *et al.* proposent, par exemple, une stratégie tout à fait originale faisant explicitement référence à la théorie du repliement des protéines (paragraphe 2.3.4) (Dong et al. 2008). Pour chaque fragment de séquence, une prédiction des structures locales probables est réalisée à l'aide de SVMs. Le chemin de repliement le plus probable est ensuite recherché par programmation dynamique en considérant des fragments non chevauchants et en optimisant globalement la probabilité de transition entre classes structurales et/ou la force de la relation structure-séquence (cf. Figure 26). Cette stratégie est nommée "*My Peeling*" par analogie à la méthode d'analyse des structures 3D en termes de noyaux compactes : le *Protein Peeling* (voir paragraphe 2.3.6) (Gelly et al. 2006a). Pour définir la bibliothèque, les auteurs utilisent d'ailleurs le *Protein Peeling* dans un premier temps afin de découper les structures en fragments non chevauchants.

L'évaluation des prédictions est le plus souvent de deux types. La méthode la plus répandue est le calcul du pourcentage de fragments pour lequel la classe prédite correspond exactement à la classe assignée à partir de la structure, *i.e.* assurant la meilleure approximation structurale (de Brevern et al. 2000; Hunter and Subramaniam 2003 ; Sander et al. 2006; Dong 2008). La deuxième évaluation correspond à la mesure de la qualité de l'approximation structurale fournie par la prédiction. Cette évaluation n'est pas triviale. Par exemple, pour Dong *et al.*, deux paramètres complémentaires sont pris en compte : la taille *w* des fragments considérés et un seuil *t* de similarité géométrique (RMSD). Ainsi, étant donnée une structure locale réelle et sa prédiction, le pourcentage de prédictions correctes est le pourcentage de résidus appartenant aux segments de longueur *w* pour lesquels une approximation plus précise que *t* est prédite (Dong et al. 2008). Dans ces conditions, si un fragment est bien prédit, tous les résidus de ce fragment sont considérés comme correctement prédits. Benros *et al.* utilisent une évaluation similaire. Toutefois, les fragments pris en compte sont chevauchants et si un fragment est bien prédit, seul le résidu central est considéré comme correctement prédits (voir paragraphe 3.3.4.1). De même, Bystroff et Baker considèrent qu'un fragment est correctement

prédit si aucun des angles dièdres prédits ne diffère de plus de 120° de ceux de la vraie structure. Pour une séquence donnée, les auteurs calculent alors le pourcentage de résidus appartenant au moins à une fenêtre de 8 résidus considérée correctement prédite (Bystroff and Baker 1998; Bystroff et al. 2000). Ces nombreuses nuances rendent les résultats publiés par les auteurs difficilement comparables directement.

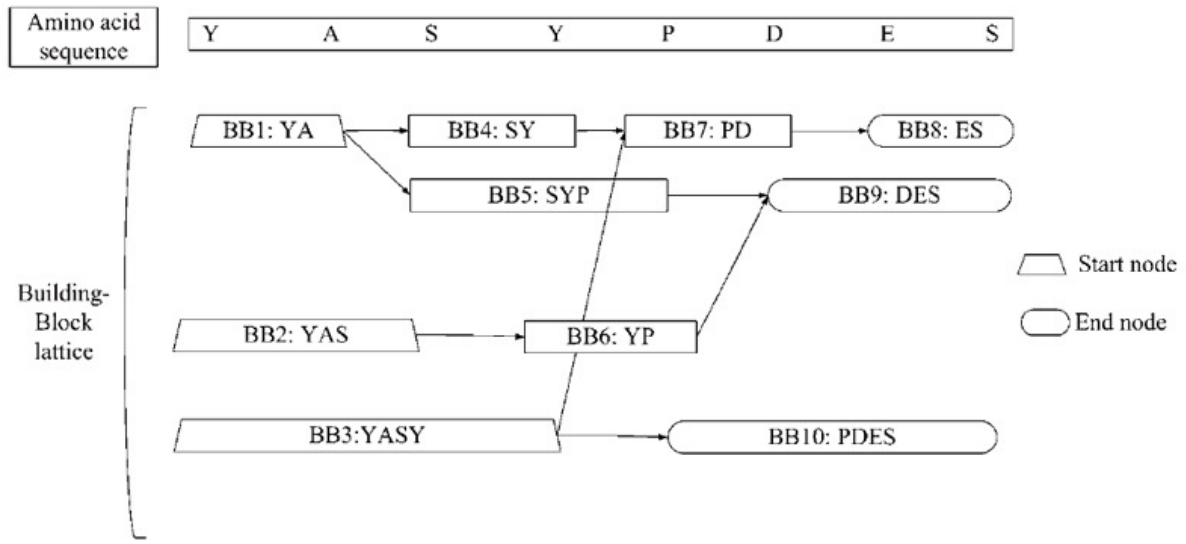

Figure 26. Illustration schématique de la stratégie de prédiction *My Peeling*.

La séquence cible à prédire est YASYPDES. La séquence et les blocs structuraux sont représentés comme par des séquences en acides aminés, en réalité, des profils sont pris en compte. De même, les blocs caractérisent en réalité des fragments de 4 à 7 résidus. Lorsqu'un segment de séquence correspond à la "*séquence*" caractéristique d'un bloc structural (*SVM*), un nœud est créé. Les arrêtes du graphe marquent alors l'adjacence des 2 blocs dans la séquence. Les blocs contenant le premier (dernier) acide aminé de la séquence sont les *blocs de Départ (d'Arrivée)*. Tous les chemins permettant d'aller d'un nœud de *Départ* à un nœud d'*Arrivée* sont un chemin de repliement possible. Les blocs finalement prédits sont ceux appartenant au chemin optimal. Figure extraite de (Dong et al. 2008).

3.1.2 Exemples de bibliothèques de fragments

3.1.2.1 Les *I-sites* de Bystroff et Baker

La bibliothèque des I-sites a été développée par Bystroff et Baker afin de prédire la structure locale à partir de la séquence en acides aminés. Sa construction est le résultat de plusieurs études (Han and Baker 1995; 1996; Bystroff and Baker 1998; Bystroff et al. 2000). Les auteurs adoptent une stratégie originale puisqu'ils choisissent de partitionner en premier lieu l'espace des séquences. Une première étape consiste en l'identification de motifs de séquence récurrents. Les structures locales associées sont analysées dans un deuxième temps.

Ainsi, des motifs de séquence récurrents non spécifiques à des familles protéiques sont tout d'abord identifiés (Bystroff and Baker 1998). Dans ce but, des alignements multiples de séquences sont construits pour 471 familles protéiques non redondantes de structures connues. Des matrices d'occurrence sont ensuite dérivées de ces alignements. Ces matrices représentent la fréquence de chaque acide aminé en chaque position des profils. Elles sont alors découpées en fragments allant de 3 à 15 positions successives. Finalement, les fragments similaires sont regroupés grâce à l'algorithme des *k-moyennes*.

Dans un second temps, les auteurs étudient la relation séquence-structure existant au sein de ces groupes. Pour chaque classe de séquences identifiée précédemment, les auteurs choisissent la structure locale adoptée le plus fréquemment par les segments comme la structure représentative et la nomme *paradigme*.

Un affinement de la relation séquence-structure est ensuite réalisé. Deux mesures de similarité structurale sont utilisées :

- La *dme* (*distance matrix error*) comparant les distances entre carbones α au sein des fragments.
- Le *mda* (*maximum deviation angle*) comparant les angles dièdres Φ et Ψ de deux structures locales. Il correspond, pour chaque position *i*, au maximum des différences observées pour deux angles Φ_{i+1} et Ψ_i.

Tous les fragments structuralement différents du *paradigme* sont exclus (*dme* > 1,4 Å et *mda* > 120°). Le profil de séquences de chaque classe structurale est ensuite recalculé et une nouvelle recherche est réalisée pour sélectionner les 400 segments les plus proches du nouveau profil. Cette étape est répétée 3 à 5 cycles jusqu'à convergence.

Finalement, une librairie de 82 groupes de motifs séquence - structure est obtenue. Chaque groupe est un *I-site*. Dans un but d'analyse, les *I-sites* de structures similaires sont regroupés et finalement 13 groupes sont présentés. Chaque classe est caractérisée par un profil de séquences et les valeurs d'angles dièdres de son *paradigme*. Deux exemples de I-sites sont présentés Figure 27.

La méthodologie employée par Bystroff et Baker pour construire la bibliothèque est originale et intéressante. Toutefois, une limitation de cette approche est de ne pas prendre en compte les motifs pour lesquelles une corrélation séquence-structure existe mais de façon tenue. Un certain nombre de séquences sont exclues durant le processus.

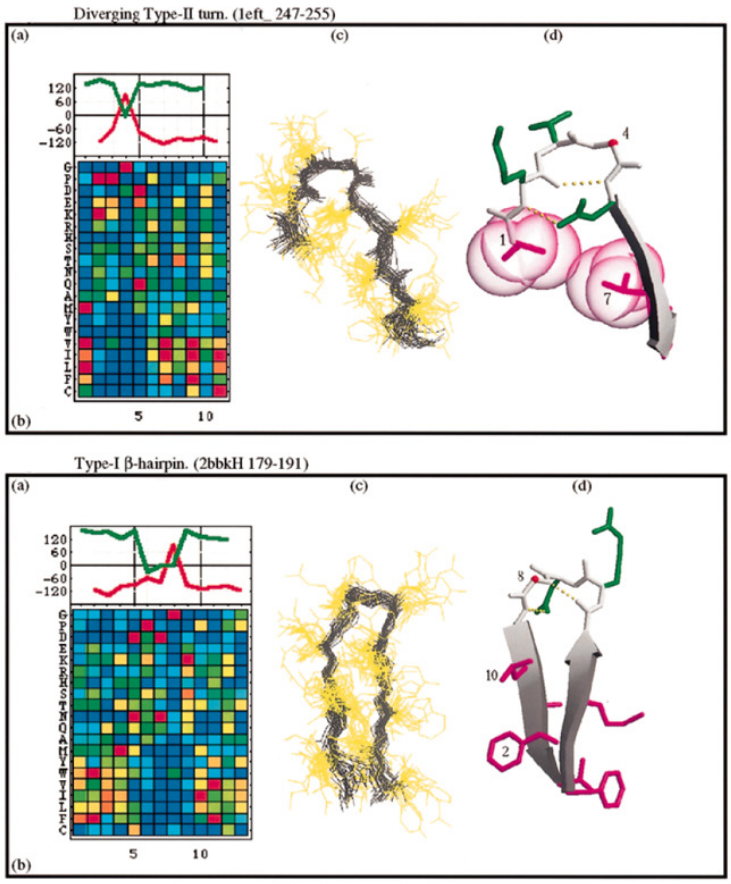

Figure 27. Exemples de *I-sites*.

(a) Angles Φ (rouge) et Ψ (vert) du *paradigme*. (b) Matrice d'occurrences en acides aminés normalisée : gradient du bleu (sous-représentations) au rouge (sur-représentations). (c) Superposition de 30 fragments membres de la classe. (d) Structure locale du *paradigme*. Figure extraite de (Bystroff and Baker 1998).

Des stratégies de prédiction ont par ailleurs été associées à cette bibliothèque des *I-sites*.

Pour une séquence cible donnée, la première méthode de prédiction des *I-sites* développée par les auteurs, réalise tout d'abord un alignement de séquences similaires pour en déduire un profil. Ce profil est ensuite découpé en fragments. Pour chaque fragment, une comparaison avec les profils des 82 *I-sites* est alors réalisée. Puis, les scores de similarités sont convertis en indices de confiance. Les fragments de la séquence cible sont ensuite ordonnés en fonction du niveau de confiance de leur prédiction. Les fragments bénéficiant de la plus grande confiance sont traités en premier. Le premier segment de liste est associé au *I-site* qui obtient le meilleur score de similarité, *i.e.,* ses angles dièdres sont attribués. Pour les segments suivants, la prédiction réalisée est prise en compte si aucun angle ne diffère de plus de 60° par rapport à des angles déjà assignés.

Cette prédiction est évaluée en utilisant des fenêtres de 8 résidus consécutifs. Un fragment est considéré comme prédit correctement si aucun des angles dièdres prédits ne diffère de plus de 120° de ceux de la structure réelle. Le taux de prédiction correct est ensuite calculé comme le pourcentage de résidus appartenant au moins à une fenêtre de 8 résidus considérée correctement prédite. Un taux de prédiction de 48 % est ainsi obtenu. Une combinaison avec une méthode de prédiction des structures secondaires (PHD (Rost et al. 1994)) permet d'atteindre 54 % de prédiction correctes.

En 2000, Bystroff et collaborateurs réactualisent la bibliothèque des *I-sites* à 180 classes de motifs séquence-structure incluant notamment des motifs plus rares. Les auteurs développent un modèle de Markov caché, HMMSTR, exploitant l'existence de transitions préférentielles entre structures locales. Ce modèle permet une amélioration de la prédiction : 59 % de prédictions correctes sont obtenues.

3.1.2.2 Les représentants structuraux de Sander et collaborateurs

Sander *et al.* se sont intéressés aux fragments structuraux chevauchants de 7 résidus de long dans le but de proposer comme Bystroff et Baker, une méthode de prédiction des structures locales. Toutefois, à l'inverse de l'approche de ces derniers, les auteurs ont partitionné l'espace des structures en lieu et place de l'espace des séquences. Leur approche est donc plus classique.

Les fragments sont décrits sous forme de matrice de distances Cα-Cα :

$$D = (d_{ij}) = \begin{bmatrix} d_{11} & d_{12} & \dots \\ d_{21} & \dots & \dots \\ \dots & \dots & \dots \end{bmatrix}$$

où d_{ij} est la distance euclidienne entre les Cα des résidus i et j du fragment d'intérêt. La comparaison entre deux fragments A et B est calculée comme suit :

$$score = \sum_{i=1}^{L} \sum_{j=1}^{L} \left| d_{ij}^{A} - d_{ij}^{B} \right|$$

Afin de discrétiser de l'espace des structures locales, les auteurs ont tout d'abord effectué une classification initiale avec l'algorithme dit du "*Leader*". Tous les fragments sont considérés de façon séquentielle et assignés à une classe. Si un fragment est *suffisamment* similaire au fragment fondateur d'une classe alors il est assigné à celle-ci, sinon il est utilisé pour fonder une nouvelle classe. Cette procédure a été conduite dix fois pour différents seuils de similarité

entre fragments. Un score de similarité seuil de 720 conduisant à la formation de 27 classes a été choisi de façon empirique pour permettre un équilibre entre un nombre limité de classes structurales et une variabilité *intra*-classe acceptable. Un raffinement des classes est finalement réalisé grâce à l'algorithme des *k-moyennes*.

La Figure 28A présente la distribution des fragments au sein des différentes classes. Une classe contient 29 % des fragments alors que les autres en contiennent 1 à 5 %. Cette classe très peuplée correspond aux structures locales hélicoïdales. En revanche, les fragments étendus (brin β) sont répartis dans différentes classes structurales et présentent presque autant de variations géométriques que les régions de boucles. Au sein des groupes, le Cα RMSD entre les fragments et leur représentant moyen est de 1,19 Å en moyennes. Des exemples de groupes structuraux sont présentés en Figure 28B.

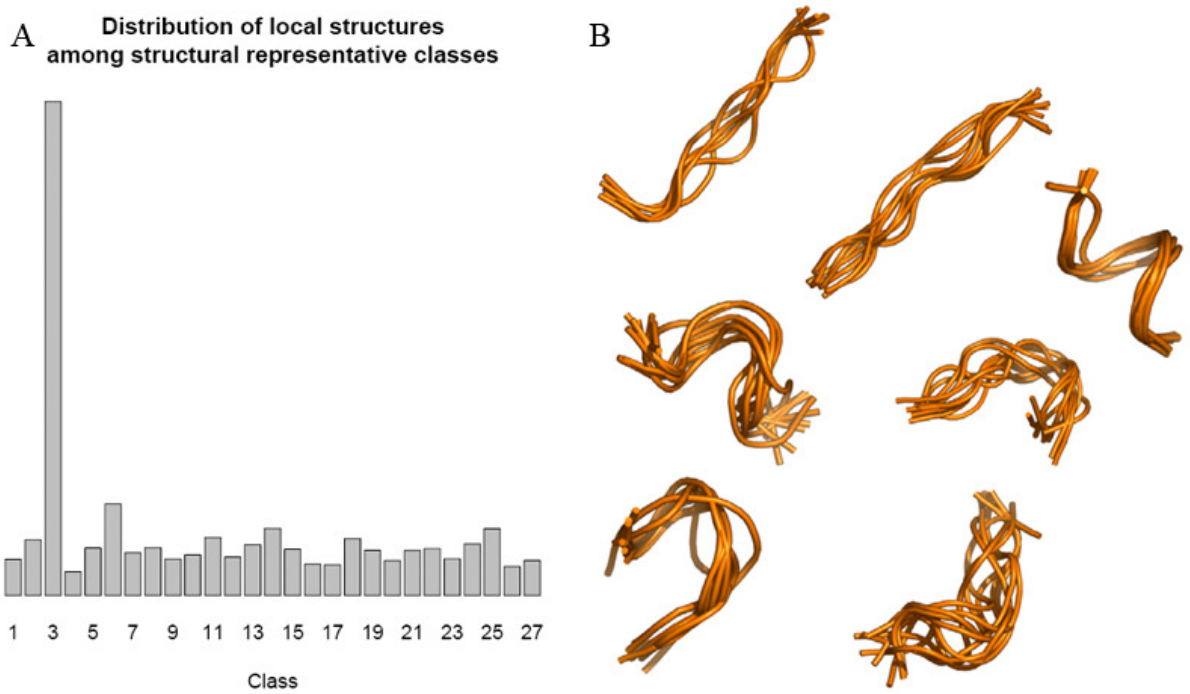

Figure 28. Les classes structurales définies par Sander et collaborateurs.

A – Proportions de fragments dans chaque classe structurale. B – Exemple de classes de structures locales. Chaque groupe est représenté par 10 fragments choisis au hasard. Figure extraite de (Sander et al. 2006).

Dans un second temps, les auteurs ont testé différentes stratégies afin de développer une méthode de prédiction des structures locales à partir de la séquence. Différentes méthodes d'apprentissage ont été testées : des arbres de décision classiques (C5), des Machines à

Vecteurs Supports (*SVM*) ou des Forêts d'arbres de décision (*Random Forest* en anglais). De même, plusieurs représentations de la séquence à prédire ont été analysées. Ainsi, les auteurs ont testé l'utilisation de la séquence seule, l'utilisation de profils évolutionnaires issus d'alignement de séquences homologues et l'utilisation d'une variante de ces profils reposant sur un codage des acides aminés en fonction de 48 propriétés physico-chimiques (*e.g.,* polarité, hydrophobicité). Les meilleurs résultats ont été obtenus en couplant les SVMs avec les profils de propriétés physico-chimiques. Un taux de prédiction de 36,15 % a été obtenu en considérant la classe assignée à partir de la structure. Le couplage des Forêts d'arbres de décision avec les profils de propriétés physico-chimiques mène à un taux de prédiction inférieur de 34,09 %. Toutefois, les auteurs choisissent ce dernier protocole car les prédictions des Forêts d'arbres sont associées à un calcul de probabilité permettant d'évaluer directement la qualité de la prédiction. La Figure 29A présente l'évolution du taux de prédiction en fonction du nombre de candidats structuraux proposés pour une séquence cible. Dans 54 % des séquences cibles, la classe structurale assignée se trouve parmi les 5 candidats structuraux les plus probables. Les courbes ROC (*Receiving Operating Characteristics* en anglais) permettant d'évaluer le taux de vrais positifs en fonction du taux de faux positifs pour chaque classe sont présentées en Figure 29B. Théoriquement, l'aire sous la courbe ROC (*Area Under the ROC curve* en anglais ou AUC) peut varier de 0,5 (pour une prédiction aléatoire) à 1 (pour une prédiction parfaite). Pour les 27 classes, l'AUC varie de 0,68 à 0,88. Les auteurs ne fournissent aucune évaluation de la précision géométrique de leurs approximations structurales.

La méthode de prédiction des structures locales développée par Sander et collaborateurs est l'une des plus récentes avec celle de (Dong et al. 2008). Les auteurs ont montré de plus qu'ils obtiennent de meilleures résultats que ceux de (Hunter and Subramaniam 2003). Ainsi, nous comparerons notamment la performance de notre méthode de prédiction des structures locales avec cette étude (voir paragraphe 4.1.4.2.1).

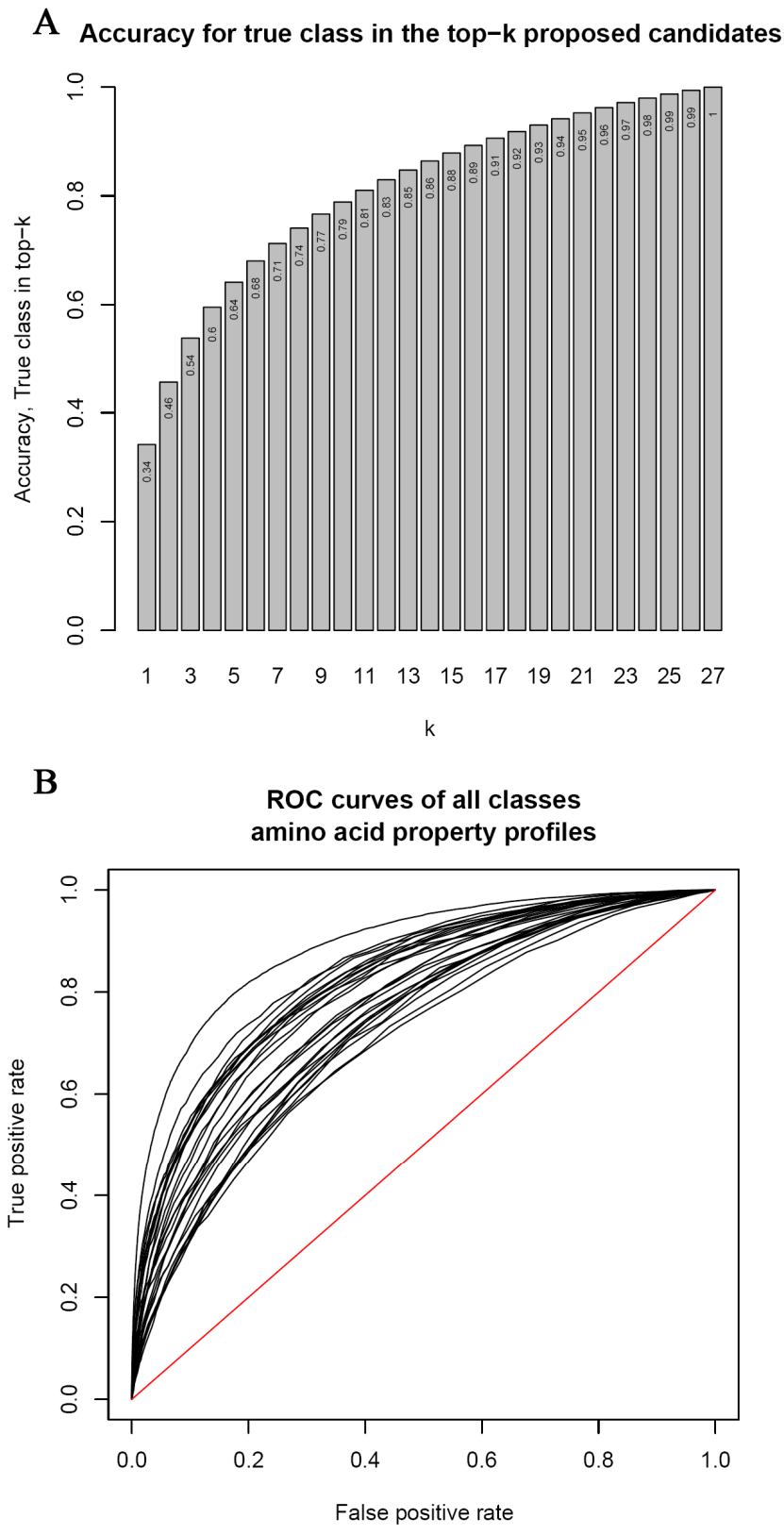

Figure 29. Résultats de Prédiction des structures locales par Sander et al.

A – Evolution du taux de prédiction en fonction du nombre candidats structuraux proposés. B – Courbes ROC pour chacune des 27 classes. Figure tirée de (Sander et al. 2006).

3.1.2.3 Blocs Protéiques et Prototypes de Structures Locales.

Au sein du laboratoire, deux alphabets complémentaires ont été développés.
L'alphabet des Blocs Protéiques (ou BPs) a été publié en l'an 2000 (de Brevern et al. 2000). Il est constitué de 16 lettres structurales de 5 résidus de long (correspondant à 8 angles dièdres). Il a été construit en considérant la similitude de séries de 8 angles dièdres et en s'appuyant sur une méthode de classification originale. De fortes relations séquence-structure ont également permis le développement de méthodes de prédiction efficaces. Les BPs sont aujourd'hui l'un des seuls alphabets mis à la disposition de la communauté scientifique. Il est également le plus utilisé à travers le monde. Aussi, cet alphabet est la base de nombreux développements méthodologiques et a été appliqué dans de nombreuses études.
En 2006, en s'appuyant sur les BPs, une seconde librairie de structure locales a été développée au sein du laboratoire pour la caractérisation de fragments plus longs de 11 résidus (Benros et al. 2006). 120 classes représentées par un Prototype de Structures Locales (PSL) ont été définies pour obtenir une approximation satisfaisante de l'espace structural. Cette longueur importante a permis de capturer des corrélations et des interactions à plus grande distance. La mise en place d'une méthode de prédiction a également été réalisée. Elle constituait pourtant un véritable challenge étant donné le nombre de classes importantes et la longueur des fragments.
Les principaux développements et résultats de mes travaux de thèse s'appuient sur le concept d'alphabet structural et plus particulièrement sur la librairie des PSLs. Je m'attacherai donc dans les deux paragraphes suivants 3.2 et 3.3 à caractériser plus en détails les BPs et les PSLs. Les méthodologies utilisées pour leur développement ainsi que les stratégies de prédiction associées seront présentées. Je mettrai également en lumière les possibilités et les limites de ces approches qui ont conduits aux développements que j'ai réalisés.

3.2 L'alphabet structural des Blocs Protéiques (BPs)

L'alphabet structural des Blocs Protéiques a été conçu pour permettre un équilibre entre *(i)* une description précise de l'ensemble des structures locales observées dans les protéines connues et *(ii)* une prédiction efficace de ces structures à partir de la séquence en acides aminés. Il a été développé en 2000 par de Brevern, Etchebest et Hazout à partir d'une base de données non redondante de 342 protéines représentatives des structures protéiques connues (de Brevern et al. 2000). En 2005, une réévaluation sur une base non redondante plus récente

de 717 protéines a confirmé sa robustesse par rapport à l'augmentation du nombre de repliements disponibles dans la PDB (de Brevern 2005; Etchebest et al. 2005).

Nous étudierons successivement sa construction, les caractéristiques principales des lettres structurales et la mise en place de la méthode de prédiction associée. Enfin, des exemples d'applications seront présentés.

3.2.1 Définition des Blocs Protéiques

L'alphabet des BPs est constitué de 16 lettres de 5 résidus de long caractérisés par des séries de 8 angles dièdres. Chaque lettre (ou BP) est la représentante d'une classe structurale rassemblant tous les fragments similaires observés dans les protéines connues.

Ainsi, les conformations de tous les fragments de 5 résidus de long de la banque de données non redondante ont été décrites en coordonnées internes par une série de huit angles dièdres caractéristiques : $V = [\psi_{n-2}, \phi_{n-1}, \psi_{n-1}, \phi_n, \psi_n, \phi_{n+1}, \psi_{n+1}, \phi_{n+2}]$, l'indice *n* étant affecté au résidu central (cf. Figure 30). Tous les fragments chevauchants sont considérés, *i.e.* une protéine de longueur *L* est donc décrite par *L-4* fragments.

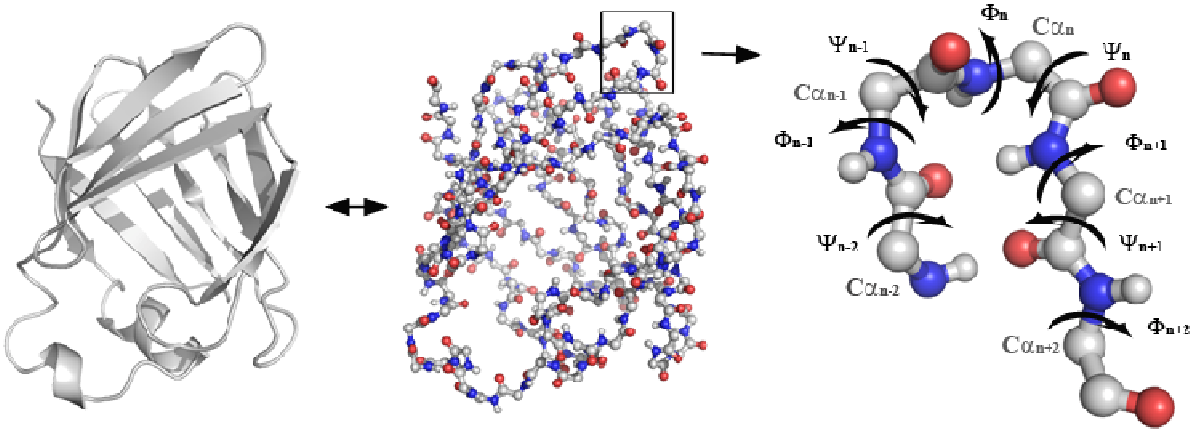

Figure 30. Définition des Blocs Protéiques : description d'un fragment de structure en une série de 8 angles dièdres.

Gauche : Représentation *cartoon* du squelette de la protéine intestinale de liaison aux acides gras de rat (code PDB 1AEL, (Hodsdon and Cistola 1997)). *Milieu* : représentation plus précise du squelette polypeptidique permettant de situer les atomes de carbone (blanc), d'azote (bleu) et d'oxygène (rouge). *Droite* : zoom sur un fragment de 5 résidus de la protéine. Les 5 Cα successifs sont notés de $C\alpha_{n-2}$ à $C\alpha_{n+2}$. Le fragment est caractérisé par la série de 8 angles dièdres $V = [\psi_{n-2}, \phi_{n-1}, \psi_{n-1}, \phi_n, \psi_n, \phi_{n+1}, \psi_{n+1}, \phi_{n+2}]$ décrivant sa conformation.

Une fois tous les fragments structuraux de la banque de données décrits, un apprentissage non supervisé est réalisé dans le but de définir un nombre limité de fragments les représentant au mieux. Ces fragments représentatifs seront nommés Blocs Protéiques. Le principe de la

méthode d'apprentissage utilisée est proche de celui des cartes de Kohonen (*Self-organized maps* ou SOMs en anglais) (Kohonen 1989; 1997). Deux étapes d'apprentissage ont été nécessaires :

Etape 1 : Le nombre de classes structurales désirées est fixé initialement à *B*. Ainsi, *B neurones* sont tout d'abord définis aléatoirement. Le terme de *neurone* fait référence à la terminologie des SOMs et correspond à un vecteur de 8 angles dièdres caractérisant un fragment tiré aléatoirement dans la banque de données. De manière itérative, chaque fragment de la banque de données est ensuite comparé à chacun de ces neurones et attribué au neurone le plus proche. Le critère utilisé est celui du RMSDA (*Root mean square Deviation on Angular Values*) (Schuchhardt et al. 1996):

$$RMSDA(V_1, V_2) = \sqrt{\frac{1}{2(M-1)} \sum_{i=1}^{i=M-1} [\psi_i(V_1) - \psi_i(V_2)]^2 + [\phi_{i+1}(V_1) - \phi_{i+1}(V_2)]^2}$$

avec V_1 et V_2 les deux vecteurs d'angles dièdres comparés.

Chaque attribution entraine une modification du neurone concerné, *i.e.,* les angles dièdres d'un neurone donné sont modifiés pour ressembler davantage au fragment qui lui est attribué. Ces modifications sont fortes pour les premiers fragments puis diminuent progressivement. Plusieurs cycles d'apprentissage sont nécessaires pour obtenir les neurones optimaux. La banque de données est donc lue plusieurs fois. A la fin de cette étape, un premier ensemble de neurones optimaux ou BPs est obtenu.

Etape 2 : Les structures protéiques de la banque de données sont alors codées en termes de BPs (cf. Figure 31). Et les fréquences de transition entre BPs successifs le long des protéines sont calculées. Une deuxième étape de raffinement est ensuite réalisée pour tenir compte des transitions préférentielles observées entre BPs. Ainsi, plusieurs cycles d'apprentissage supplémentaires sont réalisés. L'attribution d'un fragment à une classe structurale n'est plus faite uniquement en fonction du RMSDA mais également en fonction de la probabilité maximale de transition du PBs assigné au fragment précédent. Cette étape permet de tenir compte de l'architecture des protéines (par exemple, une structure correspondant à une entrée d'hélice α est suivie par une structure de cœur d'hélice α, elle-même suivie par une structure de sortie d'hélice α). Le nombre *B* de classes initialement définies est réduit progressivement durant cette dernière étape de la procédure. Si deux classes présentent une similitude de structure et de transition trop importante, la classe la moins fréquente est supprimée.

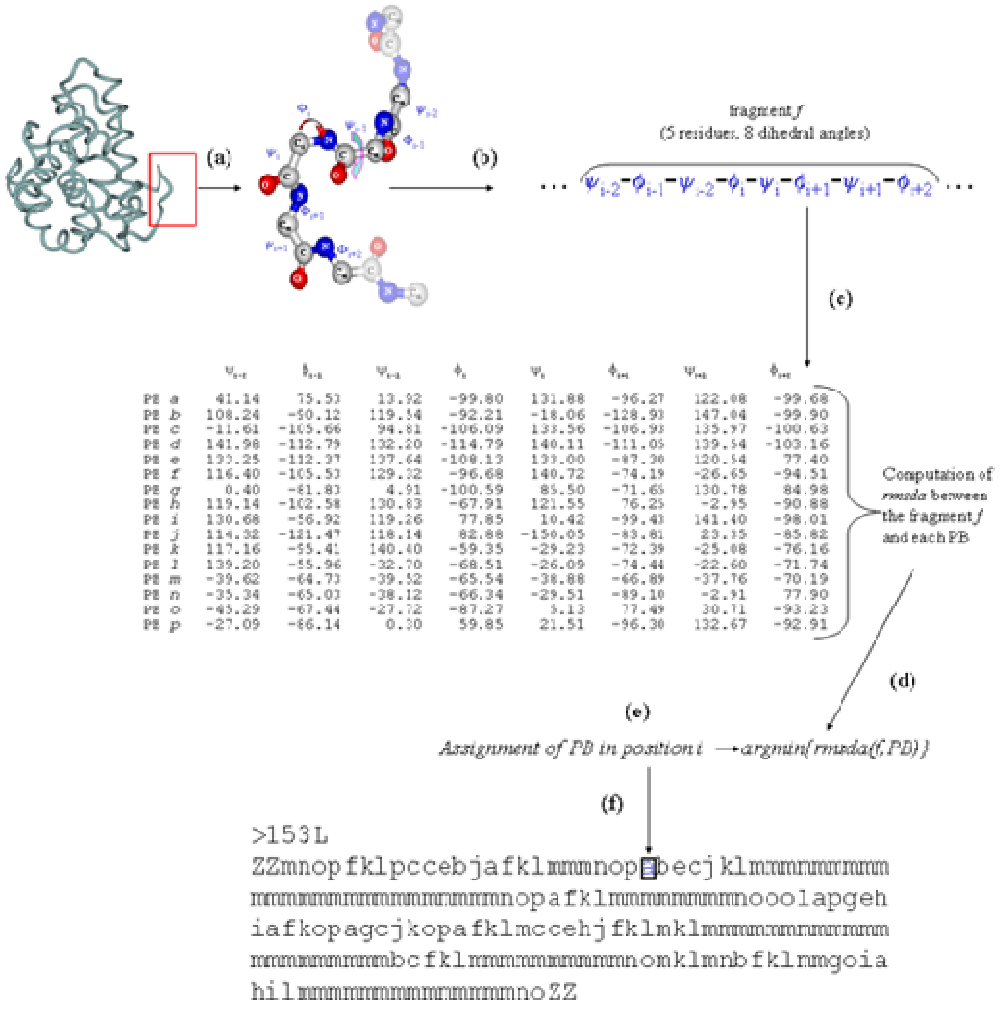

Figure 31. Codage d'une structure protéique en termes de BPs.

La protéine est le lysozyme du bactériophage T4 (code PDB 1L53 (Daopin et al. 1991)). Elle est constituée de 164 résidus. Sa structure est donc codée par 160 BPs. Les résidus ne pouvant être attribués à chaque extrémité sont notés par des Z. Figure extraite de (Etchebest et al. 2005).

Cette procédure automatique a permis à de Brevern et collaborateurs d'analyser plusieurs alphabets composés de 34 à 10 lettres. Finalement, un alphabet de 16 classes structurales a été sélectionné pour permettre un équilibre entre une description précise et des résultats de prédiction corrects (cf. paragraphe 3.2.4). Les lettres sont notées de *a* à *p*. Chaque classe est caractérisée par un BPs représentatif correspondant à une série d'angles dièdres moyens (cf. Figure 32).

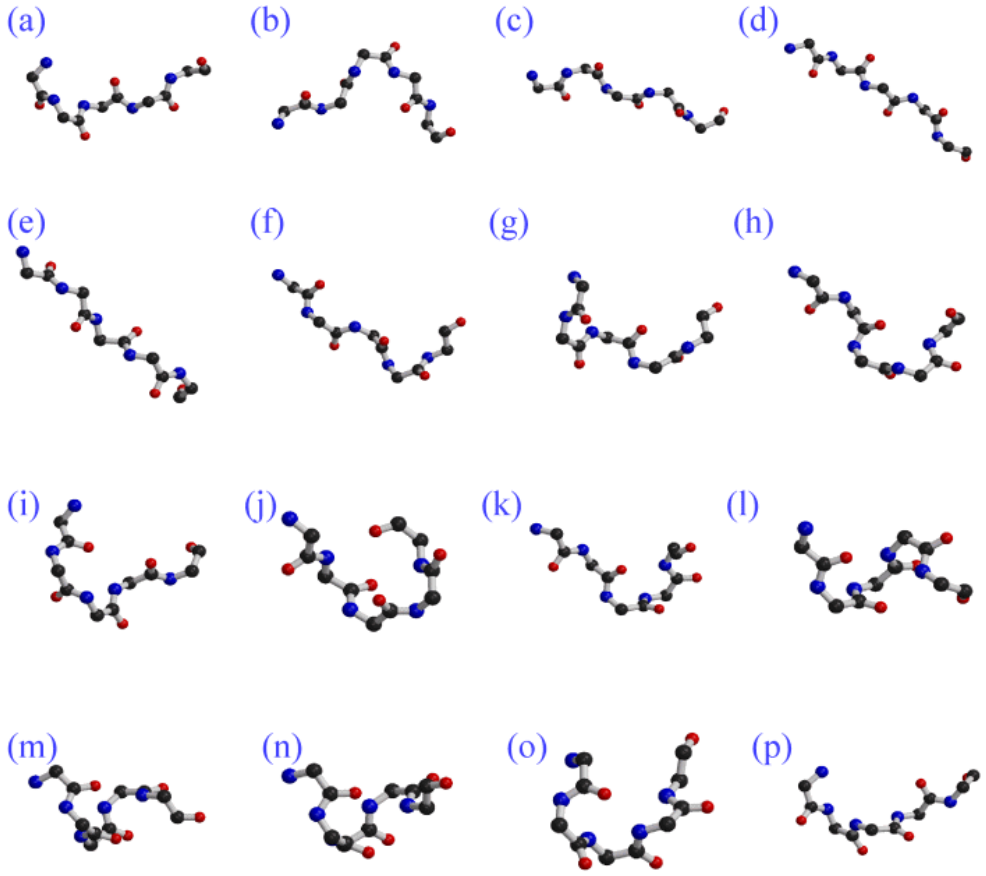

Figure 32. L'alphabet des Blocs Protéiques.

L'alphabet des 16 Blocs Protéiques est présenté. Les Blocs sont nommés par des lettres de *a* à *p*. Figure extraite de (de Brevern 2005).

3.2.2 Caractéristiques structurales des BPs

Dans un but d'analyse, les BPs peuvent être décrits en fonction de leur composition en structures secondaires et de leurs propriétés de transition (Tableau 6). Les BPs *m* et *d* correspondent respectivement aux cœurs d'hélices α et de brins β. Les blocs *a* à *c* décrivent principalement les extrémités N-terminales de brins β. Tandis que les extrémités C-terminales de ces derniers correspondent aux blocs *e* et *f*. Les BPs *g* à *j* décrivent des structures locales de boucles. Enfin, *k* à *l* et *n* à *p* correspondent respectivement aux entrées et aux sorties d'hélices α (de Brevern et al. 2000). La mesure de compacité donnée par la distance C_1-C_5 entre le $C\alpha_{n-2}$ et le $C\alpha_{n+2}$ montre un gradient allant du BP *d*, le plus étendu, aux BPs hélicoïdaux. Le bloc *n* décrivant les extrémités C-terminale des hélices est légèrement plus compact que le cœur d'hélice, *m*.

Tableau 6. Caractéristiques structurales des Blocs Protéiques.

PB	*rmsda* (°)				*rmsd* (Å)			*d*(C_1 - C_5)	*freq*	*major transitions* (%)				STRIDE		
	mean	*s.d.*	*median*	*dif.*	*mean*	*s.d.*	*median*	(Å)	(%)	1st	2nd	3rd	sum.	α (%)	coil (%)	β (%)
a	45.2	20.4	**42.2**	29.3	0.46	0.16	**0.43**	10.6	3.89	51.0 (*c*)	16.9 (*f*)	9.4 (*d*)	77.3	0.1	**75.8**	24.1
b	42.5	15.4	**41.3**	20.3	0.47	0.19	**0.43**	10.0	4.41	48.4 (*d*)	15.9 (*c*)	12.9 (*f*)	77.2	0.1	**85.3**	14.6
c	38.4	12.2	**36.3**	21.4	0.51	0.20	**0.47**	11.9	8.12	62.6 (*d*)	23.5 (*f*)	5.7 (e)	91.8	0.0	**57.6**	42.4
d	29.7	14.6	**25.7**	27.2	0.41	0.20	**0.36**	12.5	18.85	50.4 (*f*)	26.3 (*c*)	19.9 (e)	96.6	0.0	29.0	**71.0**
e	40.9	18.4	**36.2**	23.5	0.71	0.51	**0.52**	11.9	2.45	81.1 (*h*)	8.6 (*d*)		89.7	0.0	45.5	**54.5**
f	37.5	14.7	**33.7**	22.1	0.40	0.15	**0.38**	11.3	6.68	61.5 (*k*)	35.0 (*b*)		96.5	0.0	**73.3**	26.7
g	**50.6**	15.1	52.6	14.9	**0.60**	0.21	0.62	9.5	1.15	37.5 (*h*)	29.6 (*c*)	16.1 (*o*)	83.2	13.3	**80.2**	6.4
h	**47.0**	17.9	50.0	20.9	0.46	0.18	**0.42**	8.5	2.40	68.0 (*i*)	13.8 (*j*)	8.5 (*k*)	90.3	2.0	**76.2**	21.9
i	**43.4**	18.6	44.3	25.0	0.41	0.20	**0.37**	8.6	1.86	82.8 (a)	6.2 (*l*)		89.0	2.0	**90.3**	7.7
j	49.0	18.7	**48.2**	19.6	0.83	0.48	**0.76**	8.4	0.83	21.7 (*b*)	14.8 (a)	14.7 (*k*)	51.2	8.0	**81.6**	10.4
k	35.9	17.0	**31.8**	25.4	0.30	0.12	**0.27**	7.5	5.45	77.2 (*l*)	10.5 (*b*)	6.2 (*o*)	93.9	49.3	**50.2**	0.5
l	32.5	20.8	**23.6**	27.3	0.53	0.29	**0.46**	7.3	5.46	68.2 (*m*)	8.6 (p)	7.1 (*c*)	83.9	**61.0**	38.6	0.4
m	15.0	16.2	**7.6**	40.1	0.31	0.25	**0.21**	6.6	30.22	34.9 (*n*)	15.7 (*p*)	11.3 (*k*)	61.9	**92.3**	7.6	0.1
n	26.8	22.3	**15.0**	31.2	0.31	0.22	**0.23**	6.5	1.99	92.4 (*o*)			92.4	**75.7**	24.0	0.3
o	**38.3**	23.1	40.3	27.1	0.48	0.24	**0.43**	6.9	2.77	78.2 (*p*)	6.5 (*m*)	5.6 (*i*)	90.3	**50.8**	49.0	0.2
p	**43.8**	20.6	52.6	25.9	0.47	0.20	**0.43**	9.4	3.47	58.6 (a)	23.7 (*c*)	7.6 (*m*)	89.9	17.1	**81.3**	1.6
Mean	30.1	20.1	26.1	29.5	0.41	0.25	0.34	9.2	100.0	63.9	10.8	7.3	89.3	37.8	39.7	22.5

Pour chaque bloc de *a* à *p* sont donnés : *(i)* la moyenne (*mean*), l'écart-type (*s.d.*) et la médiane (*median*) du RMSDA calculé au sein des classes structurales. *dif.* correspond à la différence entre le RMSDA correspondant au meilleur BP et le RMSDA du deuxième plus proche BP. *(ii)* la moyenne (*mean*), l'écart-type (*s.d.*) et la médiane (*median*) du RMSD calculé au sein des classes structurales. *(iii)* La distance entre les extrémités du BPs ($d(C_1\text{-}C_5)$). *(iv)* la fréquence de chaque bloc (*freq*), *(v)* les trois transitions les plus fréquentes vers d'autres blocs (*major transitions*) et la somme de ces transitions (*sum*), seules les transitions de plus de 5 % sont notées. *(vi)* La répartition de l'assignation en structures secondaires du résidu central par STRIDE (α-helix, coil and β-strand), les fréquences de plus de 50% sont mises en valeur en gras. Tableau adapté de (de Brevern 2005).

Les 16 BPs permettent une approximation angulaire moyenne de 30 degrés (σ = 20°) de l'ensemble des fragments des protéines de la banque de données (de Brevern 2005) (Tableau 6). Les variabilités internes les plus faibles sont observées pour les blocs fortement hélicoïdaux *m* et *n*, les plus fortes concernent les blocs dérivant les boucles. Afin d'évaluer le pouvoir discriminatif des BPs, la différence entre l'approximation donnée par le BP le plus proche d'un fragment donné et l'approximation donnée par le deuxième BP le plus proche. Cette différence est importante : 29,5° en moyenne. Ainsi, si le meilleur BP donne une approximation de 30° comme nous l'avons vu, le BP du second rang donne une approximation bien moins bonne de 59,5° en moyenne. Les PBs caractérisent donc de manière assez spécifique les structures locales connues.

3.2.3 Spécificités de séquence des BPs

Une fois les structures protéiques codées en BPs, une matrice d'occurrences en acides aminés peut être calculée pour chaque BP. En effet, un BP donné est associé à *N* fragments de séquences en acides aminés. Ainsi, en chaque position, il est possible de calculer la fréquence

observée de chacun des 20 acides aminés. Pour prendre en compte l'information apportée par les résidus environnants, la fenêtre de séquence considérée est élargie de 5 résidus de part et d'autre du BP, une fenêtre de 15 résidus est donc prise en compte (de Brevern et al. 2000; Etchebest et al. 2005).

Afin de mettre en évidence les spécificités de séquence de chaque BP par rapport à la distribution globale dans la banque de données, les matrices d'occurrences peuvent être normalisées en termes de Z-scores :

$$Z(AA_k^j) = \frac{n_{obs}(AA_k^j) - n_{th}(AA_k)}{\sqrt{n_{th}(AA_k)}} \quad \text{avec} \quad n_{th}(AA_k) = N_x \cdot F(AA_k)$$

$n_{obs}(AA_k^j)$ est le nombre d'occurrences de l'acide aminé de type k (de 1 à 20) en position j de la fenêtre de séquence. $n_{th}(AA_k^j)$ est le nombre théorique (ou attendu) d'occurrences de l'acide aminé AA_k en position j. $F(AA_k)$ est la fréquence de l'acide aminé AA_k dans la banque de données. N_x est le nombre de séquences associées au BP x d'intérêt. Les Z-scores positifs et supérieurs à un seuil fixé ε correspondent à des sur-représentations significatives d'acides aminés. En revanche, les Z-scores négatifs et inférieurs au seuil –ε correspondent à des sous-représentations significatives. Deux valeurs seuil ont été souvent utilisées, ε = 1,96 et ε = 4,4, correspondant respectivement à un risque de première espèce (risque α) de 5.10^{-2} et de 10^{-5}.

Par ailleurs, la mesure de divergence asymétrique de Kullback-Leibler (ou *KLd*) permet d'analyser l'informativité de chaque position (Kullback and Leibler 1951). Il permet de mesurer la dissimilarité entre la distribution des acides aminés en chaque position pour chaque BP et la distribution des acides aminés dans la banque de données. Le score du KLd en position j est calculé comme suit :

$$KLd(j) = \sum_{k=1}^{k=20} p_{obs}(AA_k^j) \ln\left[\frac{p_{obs}(AA_k^j)}{p_{th}(AA_k)}\right]$$

$p_{obs}(AA_k^j)$ est la probabilité d'observer l'acide aminé AA_k en position j de la fenêtre de séquence. $p_{th}(AA_k)$ est la probabilité d'observer l'acide aminé AA_k. La significativité des valeurs de $KLd_s(y)$ est évaluée par un test du χ^2. En effet, le produit ($2N_x$ x $KLd_x(j)$) suit une loi du χ^2 à 19 degrés de liberté. Un seuil de significativité des valeurs de *KLd* est donc déterminé pour chacune des matrices d'occurrences en acides aminés, correspondant au rapport : (χ^2 / $2N_x$), pour un risque de première espèce α donné.

Ainsi, de fortes relations séquence-structure ont pu être mises en évidence au sein des blocs protéiques. Les positions les plus informatives correspondent aux 5 résidus centraux directement associés aux BPs. Les distributions d'acides aminés observées dans le BPs *m* et *d* sont proches de celles observées pour les hélices α (L, A, M, Q, E) et les feuillets β (I, V, F, Y) respectivement. Les préférences des blocs *k, l, n, o*, correspondant aux sorties d'hélices et de brins, impliquent les acides aminés S, N, D, E, P et Q. Les BPs associés aux structures non répétitives présentent de plus des spécificités de séquences significatives. La glycine G est par exemple sur-représentée au centre du BP *j*. Les acides aminés N et P sont également sur-représentées au centre du BP *h*.

3.2.4 Prédictions des BPs à partir de la séquence

Plusieurs méthodes de prédiction des BPs à partir de la séquence ont été proposées à ce jour. La méthode de prédiction initialement développée était basée sur une stratégie Bayésienne (de Brevern et al. 2000). Elle reposait sur le calcul de la probabilité d'observer le BP *x* connaissant la séquence cible. Cette probabilité est calculée grâce au théorème de Bayes et à l'utilisation des matrices d'occurrences associées aux BPs (cf. Figure 33). Cette approche permettait d'obtenir 34,4% de prédictions correctes. Une bonne prédiction est enregistrée lorsque le BP prédit (celui ayant obtenu le meilleur score R_x (voir Figure 33)) correspond au BP assigné. Il est important de noter que l'évaluation de la performance de la méthode a été réalisée sur un jeu de protéines n'ayant pas servi à développer l'alphabet.

Un premier raffinement de la méthode a permis une amélioration de 6,3% du taux de prédiction (soit un taux de 40,7%). La stratégie utilisée reposait sur le concept des *Familles Séquentielles* selon lequel plusieurs types de séquences peuvent être associés à un même repliement. Un BP peut donc être associé à plusieurs matrices d'occurrences. Ainsi, pour chaque BP, *(i)* le nombre de types de séquences et *(ii)* la répartition des séquences au sein de chaque type sont optimisés à l'aide d'une stratégie proche des Cartes Topologiques de Kohonen (de Brevern et al. 2000).

En 2005, une nouvelle méthode a été proposée et évaluée (Etchebest et al. 2005). Celle-ci repose sur le concept des *Familles Séquentielles* mais utilise une stratégie plus sophistiquée pour leur construction. L'apprentissage des *Familles* se déroule en deux étapes : (1) une première étape similaire à la stratégie utilisée précédemment et (2) une seconde étape de raffinement basée sur une approche proche du recuit simulé. L'objectif de cette seconde étape était d'optimiser les *Familles* de chaque PB en fonction du taux de prédiction global.

Finalement, trois BPs ont été divisés en deux *Familles* (*b, c, f*). De plus, les BPs *d* et *m* ont été divisés en 3 et 6 *Familles* respectivement (cf. Figure 34). Ce découpage des classes structurales en sous groupes de séquences plus spécifiques a permis de renforcer la relation séquence-structure. Une augmentation de 8% a été obtenue, soit un taux de prédictions correctes de 48,7%.

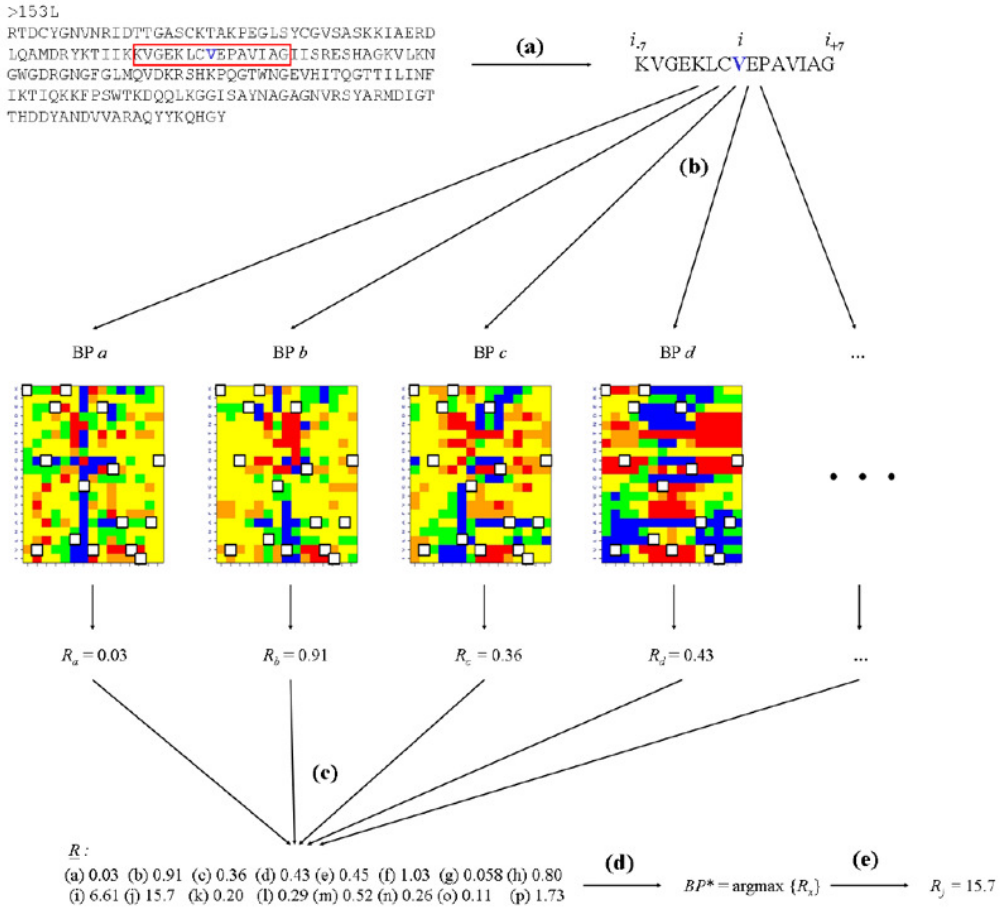

Figure 33. Prédiction Bayésienne simple des BPs à partir de la séquence.

(a) La prédiction est réalisée pour chaque fenêtre de séquence de la protéine cible. Notons qu'une fenêtre de séquence est longue de 15 résidus, les positions sont donc numérotées de *i*-7 à *i*+7, *i* étant le résidu central. *(b)* Le théorème de Bayes permet de calculer un score R_x pour chaque BP *x*. Ce score est le logarithme népérien du rapport entre la probabilité d'observer le BP *x* sachant la séquence et la probabilité d'observer le BP *x* dans la banque de données. Ce score repose sur l'utilisation des matrices d'occurrences en acides aminés. Pour chaque matrice, les acides aminés sont en ordonnée dans l'ordre suivant [I, V, L, M, A, F, Y, W, C, Q, P, G, H, S, T, N, D, E, R, K] et la position dans la fenêtre de séquence est en abscisse. Les valeurs du Z-score sont décrites par les couleurs : *(bleu)* Z-score < -4,4 (sous-représentation), *(vert)* -4,4 ≤ Z-score < - 1,96, *(jaune)* -1,96 ≤ Z-score < 1,96, *(orange)* 1,96 ≤ Z-score ≤ 4,4, *(rouge)* Z-score > 4,4 (sur-représentation). (c) Les 16 scores R_x sont comparés. Le BP prédit est celui ayant obtenu le score maximale. (e) Ici, le BP *j* est prédit. Figure extraite de (Etchebest et al. 2005).

En revanche, le couplage des *Familles Séquentielles* avec une prédiction des structures secondaires (PSIPRED) n'a pas permis d'améliorer significativement la prédiction. En effet, un gain de 1 % seulement a été obtenu et un biais favorisant la prédiction des BPs répétitifs au détriment des autres a été introduit.

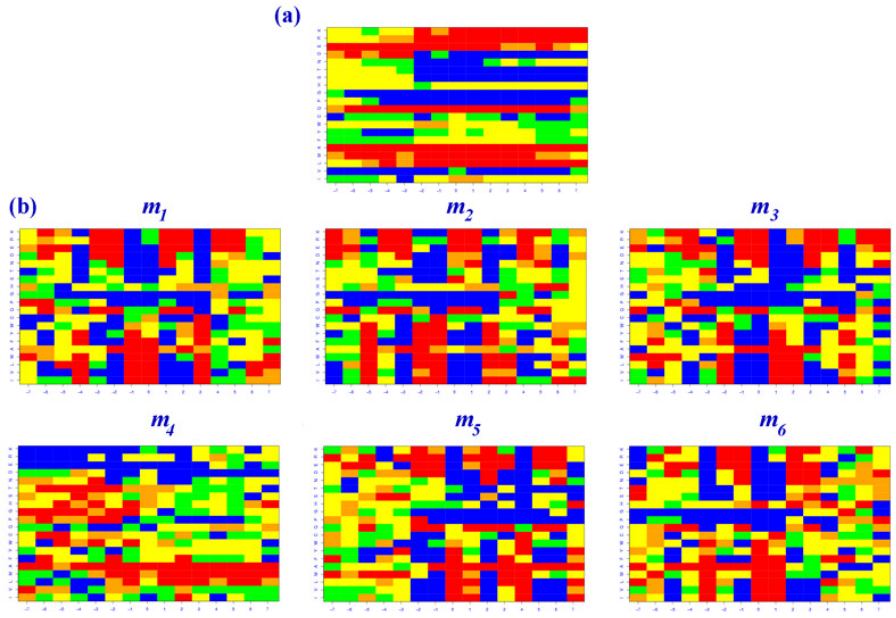

Figure 34. Les Familles Séquentielles du BP m.

Les matrices de Z-scores associées (a) au BP *m* et (b) à ses familles séquentielles sont présentées. Figure extraite de (Etchebest et al. 2005).

Récemment, Zimmermann et Hansmann ont développé une nouvelle méthode de prédiction des BPs : LOCUSTRA (Zimmermann and Hansmann 2008). Leur stratégie repose sur *(i)* l'enrichissement des séquences par des données évolutionnaires et sur *(ii)* l'utilisation de Machines à Vecteurs Supports (ou *Support Vector Machines* en anglais, SVMs). Ainsi, la prédiction ne repose plus seulement sur la séquence seule mais sur un profil statistique issu de l'alignement de séquences similaires à la séquence cible. Les séquences similaires sont automatiquement recherchées dans une banque de données de séquences grâce au programme PSIBLAST (Altschul et al. 1997). Nous reviendrons sur la collecte des données évolutionnaires, sur la création de profils statistiques à partir d'alignements et sur le principe des SVMs dans le paragraphe 4.1.2.3. J'ai en effet également utilisé ces techniques dans le cadre des développements méthodologiques que j'ai réalisés. Cette approche a permis à Zimmermann et Hansmann d'obtenir un excellent taux de prédiction de 61,0% de prédictions correctes pour 16 classes.

Par ailleurs, dès 2000, de Brevern et collaborateurs proposent l'élaboration d'un indice de confiance permettant d'évaluer directement la qualité de la prédiction. Ainsi, pour évaluer l'incertitude liée à la prédiction, les auteurs calculent l'entropie *H* liée à la distribution des scores R_x. En effet, une forte homogénéité des scores R_x pour une séquence cible donnée démontre une faible spécificité de séquence et devrait mener à une prédiction peu précise. A l'inverse, un score élevé au premier rang devrait mener à une prédiction correcte. *H* est calculée comme suit :

$H = -\sum_x S_x \ln(S_x)$ avec $S_x = R_x / \sum R_x$, x est l'indice correspondant aux différents BPs.

Cette entropie H est ensuite transformée en $N_{eq} = exp(H)$ (Nombre équivalent). Cette mesure varie de 1 (lorsqu'un BP unique est clairement prédit) à *M* (lorsque *M* BPs sont prédits de manière équiprobable). Cet indice permet de déterminer au mieux le nombre de BPs à prédire pour une séquence cible. Par ailleurs, le N_{eq} est lié au taux de prédiction. Il permet donc également de sélectionner les sites plus faciles à prédire et garantissant un taux de prédiction minimal. En 2005, une prédiction du taux de prédiction à partir du N_{eq} est proposée (Etchebest et al. 2005). Elle repose sur régression linéaire et permet une très bonne approximation avec seulement une erreur moyenne de 5%.

3.2.5 Applications des BPs et travaux dérivés

L'alphabet des BPs a été développé pour décrire les conformations locales adoptées par le squelette polypeptidique (de Brevern 2005) mais également pour la prédiction des structures locales (de Brevern et al. 2000; de Brevern et al. 2004; Etchebest et al. 2005; de Brevern et al. 2007).

Depuis, l'alphabet a été utilisé dans d'autres laboratoires pour la reconstruction de structures protéiques globulaires (Dong et al. 2007), pour le développement de peptides (Thomas et al. 2006), la caractérisation de sites de liaison (Dudev and Lim 2007) (voir Figure 35) ainsi que pour le développement d'un potentiel statistique local (Li et al. 2009). Par ailleurs, les BPs ont été comparés par Karchin et collaborateurs avec huit autres alphabets structuraux (Karchin et al. 2003). Cette étude montre clairement que les BPs sont hautement informatifs et sont les plus prédictibles parmi ceux testés. Ainsi, l'alphabet des BPs est aujourd'hui devenu le plus utilisé par la communauté scientifique internationale.

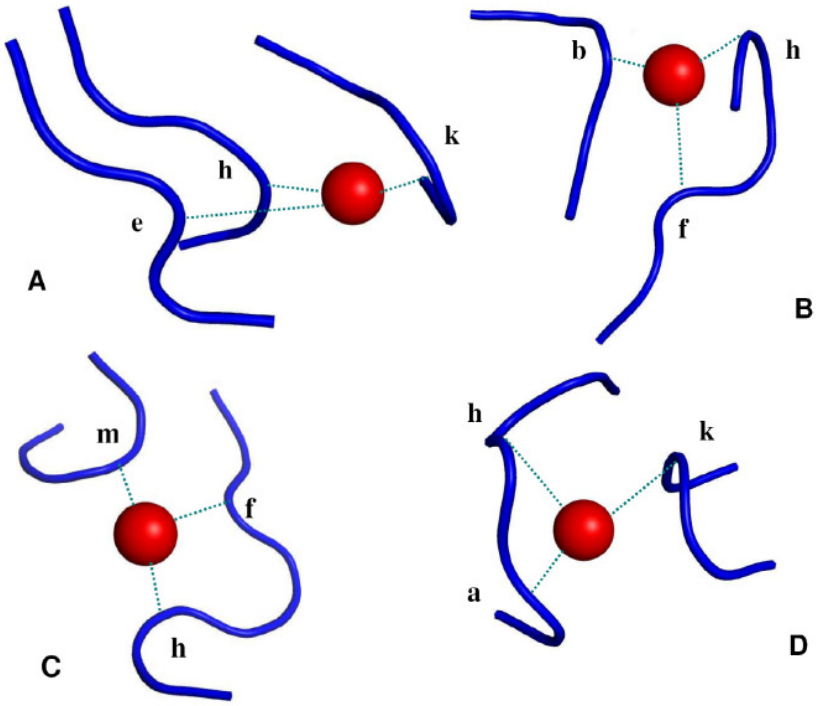

Figure 35. Les quatre motifs structuraux conservés caractérisant les sites de liaison des protéines au Magnésium.

Les protéines liant le magnésium ne partagent pas fréquemment d'identité de séquence au niveau de leur site de liaison. Cependant, certaines présentent un site de liaison de structure similaire. 4 motifs structuraux ont été identifiés en utilisant les BPs. *(a)* Motif *e*(24–47)*h*(24)*k* observé chez des lyases et des isomérases, les chiffres entre parenthèse indiquent le nombre de résidus entre les BPs. *(b)* Motif *f*(1)*h*(109–349)*b* observé au sein d'isomérases et d'hydrolases. *(c)* Motif *f*(2)*h*(126–158)*m* observé dans des structures d'hydrolases. (d) Motif *k*(26–29)*h*(1)*a* observé dans des transférases, des oxydo-réductases et des lyases. Figure extraite de (Dudev and Lim 2007).

Au sein du laboratoire, les BPs ont notamment été utilisés pour l'analyse des contacts au sein des protéines (Faure et al. 2009) (paragraphe 2.3.5), pour la comparaison/superposition des structures protéiques (Tyagi et al. 2006a; Tyagi et al. 2006b; Tyagi et al. 2008) ou encore pour la modélisation d'une protéine transmembranaire (de Brevern et al. 2005).

Dans les paragraphes suivants, je développerai plus particulièrement trois applications. Je présenterai tout d'abord deux études auxquelles j'ai pu participer : la construction d'un alphabet réduit d'acides aminés pour l'analyse des mutations (Etchebest et al. 2007) et le développement d'une méthode de prédiction des boucles courtes basée sur les BPs (Tyagi et al. 2009a). Dans un second temps, je présenterai les travaux fondateurs de de Brevern et collaborateurs pour la description et la prédiction de long fragments de structures (de Brevern et al. 2002; de Brevern et al. 2007). Ces travaux me permettront d'introduire les développements de Benros et associés (paragraphe 3.3) (Benros et al. 2006) sur lesquels reposent mes principaux travaux de thèse.

3.2.5.1 Analyse et prédiction des boucles courtes (Article 7)

Comme nous l'avons vu (paragraphe 2.3.8.1), la prédiction de la structure tridimensionnelle des boucles est un champ de recherche encore très actif aujourd'hui. Les méthodes récentes utilisent des stratégies sophistiquées de modélisation *ab initio* impliquant un large échantillonnage de structures locales et des champs de force. Elles s'appuient, de plus, sur la connaissance des extrémités des structures secondaires de part et d'autres des boucles à modéliser (Fiser et al. 2000; Xiang et al. 2002; de Bakker et al. 2003; Zhu et al. 2006; Soto et al. 2008).

Dans une étude précédente, les performances de la prédiction des BPs pour la prédiction des boucles courtes de 2 à 6 résidus avaient été évaluées (Fourrier et al. 2004). Ces boucles sont les plus fréquemment observées. La méthode de prédiction des BPs utilisée est la méthode initiale publiée dans (de Brevern et al. 2000). En effet, la méthode plus sophistiquée publiée dans (Etchebest et al. 2005) et basée sur les *Familles séquentielles* n'était pas applicable ici, le nombre de séquences disponibles pour l'apprentissage étant trop faible. Un apprentissage dédié aux boucles avait été réalisé et 41,2 % de prédictions correctes avaient été obtenues.

Récemment, au cours du travail de thèse de Manoj Tyagi, nous avons étendu cette analyse en nous plaçant dans le cadre de la modélisation par homologie (Tyagi et al. 2009b) (Article 7). En effet, lors de la modélisation d'une protéine, les boucles sont modélisées au cours des dernières étapes après le cœur et les structures secondaires régulières (paragraphe 2.3.8.1). Les extrémités des boucles sont donc connues. Ainsi, nous avons évalué la prédiction des BPs effectuées avec la connaissance *a priori* de la structure de leurs extrémités. Quatre catégories de boucles ont été définies : les boucles αα (bordée par une série de BPs *mm* à chacune de leurs extrémités), les boucles αβ (bordé par une série *mm* en N-terminale et une série *dd* en C-terminale), les boucles βα et les boucles ββ. Pour chaque catégorie de boucle, un apprentissage spécifique a été effectué. La Figure 36 permet de comparer les résultats de prédiction avec et sans connaissance *a priori* des extrémités des boucles. Le taux de prédiction des boucles sans aucune connaissance des extrémités est de 43,2 %[5]. Les taux de prédiction des boucles αα, αβ, βα et ββ, connaissant leurs extrémités, sont respectivement de 56,0 %, 62,94 %, 43,9 % et 49,6 %. Ainsi, tous les types de boucles bénéficient d'une augmentation du nombre de prédictions correctes de plus de 6 % à l'exception des boucles βα. L'évaluation de la prédiction non plus en termes de classes structurales (*BP prédit = BP*

[5] Ce taux est légèrement différent de celui trouvé dans Fourrier, L., Benros, C., and de Brevern, A.G. 2004. Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics* **5:** 58. car l'évaluation a été réalisée sur une banque de structures protéiques réactualisée.

assigné) mais en termes d'approximation est également améliorée. Une approximation meilleure que 2 Å est obtenue dans 66 % des cas. La connaissance des extrémités permet d'augmenter ce chiffre à 68% pour les boucles βα, 74% pour les ββ, 76% pour les αα et 82% pour les αβ.

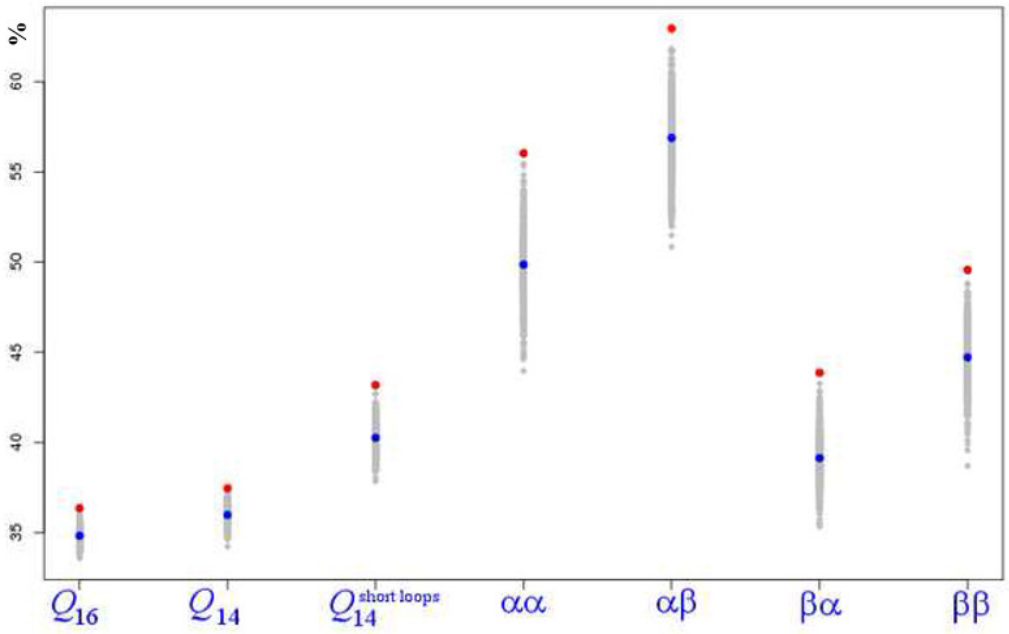

Figure 36. Distributions des taux de prédiction.

Sont présentés les taux de prédiction correctes obtenus : pour la prédiction classique de 16 BPs (Q_{16}), pour la prédiction des boucles uniquement, les BPs *m* et *d* étant exclus (Q_{14}), pour la prédiction des boucles courtes ($Q_{14}^{\text{short loops}}$), et pour la prédiction des boucles αα, αβ, βα, ββ. Les points gris correspondent aux résultats obtenus lors de 1000 cycles de validation croisée. Les points bleus correspondent aux valeurs moyennes et sont celles auxquelles se réfère le texte. Les points rouges correspondent aux valeurs maximales obtenues. Figure extraite de (Tyagi et al. 2009b).

Finalement, cette étude démontre la puissance des BPs dans le cadre de la prédiction des boucles. L'utilisation de notre méthode de prédiction pourrait s'avérer très pertinente pour la modélisation de boucles *ab initio* car elle permettrait de limiter très rapidement l'espace de recherche conformationnelle en fonction des spécificités de séquence.

3.2.5.2 Etude des conséquences structurales de mutations (Article 8)

Comme nous l'avons vu au paragraphe 3.2.4, le concept des *Familles séquentielles* selon lequel plusieurs séquences peuvent adopter un repliement similaire, permet une amélioration significative de la prédiction des BPs en renforçant la relation séquence-structure. Ainsi, le monde des séquences est plus vaste que le monde des structures.

Des études expérimentales et théoriques ont suggéré que la complexité des séquences protéiques dans son intégralité n'était pas nécessaire pour un repliement correct des protéines

(Clarke 1995; Kuhlman and Baker 2004). En conséquence, un regroupement des acides aminés en fonction de leurs propriétés permettrait de simplifier l'univers des séquences et serait un outil de choix pour le domaine de l'ingénierie des protéines. Ainsi, différents travaux ont été menés dans le but de développer un alphabet d'acides aminés minimal. L'alphabet le plus simple décrit seulement deux états : hydrophobique et polaire. Il a été utilisé pour le développement de structures stables présentant des hélices α et des feuillets β (Wei et al. 2003; Hecht et al. 2004). Quelques études expérimentales ont été basées sur des alphabets d'acides aminés réduits plus importants. Riddle et collaborateurs ont conçu un domaine SH3 de 57 acides aminés grâce à un alphabet de 5 lettres (I, A, G, E et K) (Riddle et al. 1997). Akanuma *et al.* ont plus récemment remplacé 88% des résidus de l'ororate phosphoribosyltransferase d'*Escherichia coli* en utilisant un alphabet de 9 lettres (A, D, G, L, P, R, T, V et Y) (Akanuma et al. 2002). Différentes études théoriques ont également été réalisées en se basant soit sur les structures protéiques, soit sur les séquences. Certaines analyses reposent sur les fréquences de contacts entre deux résidus (Wang and Wang 1999), d'autres sur la distribution des acides aminés en fonction les structures secondaires (Liu et al. 2003). Par ailleurs, des études privilégiant l'analyse des séquences ont exploité l'influence des résidus voisins (Rogov and Nekrasov 2001) ou des alignements de séquences (Li et al. 2003). Les résultats montrent que ces études ne parviennent pas à un consensus clair et que la problématique n'est pas triviale.

Dans une étude récente, en nous appuyant sur les BPs, nous avons étudié la distribution des acides aminés et leurs équivalences dans un contexte structural (Etchebest et al. 2007) (Article 8). La distribution des acides aminés au sein des BPs a été exploitée pour définir des groupes d'acides aminés équivalents par classification hiérarchique. Cette analyse a été réalisée tout d'abord localement par BP, puis globalement, pour tous les PBs simultanément. L'analyse globale montre que la glycine et la proline présentent des spécificités très importantes car ils sont associés à des conformations particulières du squelette polypeptidique. Les 18 acides aminés restants sont regroupés en 3 groupes : 1 - (I,V,F,Y,W), 2 - (A,L,M,E,Q,R,K), 3 - (N,D,H,S,T,C) (cf. Figure 37). Au sein de ces groupes, des comportements spécifiques vis-à-vis des structures locales peuvent également encore être observés. Au sein du groupe 1, les aliphatiques forment un groupe plus proche que les aromatiques. Dans le groupe 2, les hydrophobes peuvent être rassemblés d'une part et les polaires ou chargés avec une longue chaîne d'autre part. Enfin, dans le groupe 3, l'asparagine et l'aspartate présentent des préférences plus proches entre eux qu'avec le sous-groupe H, S, T et C. De manière

intéressante, les préférences des acides aminés vis-à-vis des structures locales ne dépendent pas uniquement de leurs propriétés physico-chimiques. En effet, la leucine et l'isoleucine ne différant que par un groupe CH2, ne sont pas classées dans le même groupe. De même, le glutamate et l'aspartate, tous deux chargés négativement, sont dans des classes différentes.

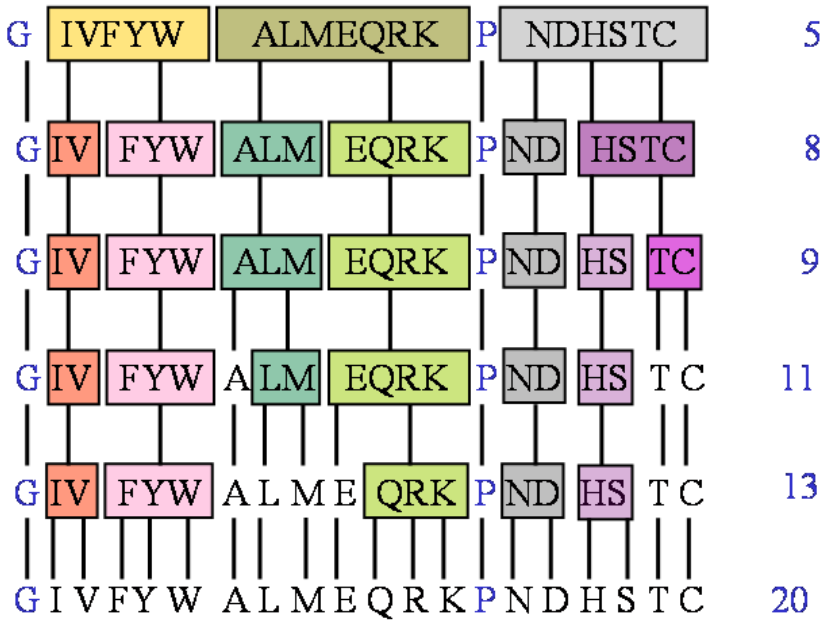

Figure 37. Les différents groupes d'acides aminés mis en évidence par notre étude.

Figure extraite de (Etchebest et al. 2007).

Localement, les associations entre acides aminés sont sensibles au BP étudié. Cependant, les groupes restent relativement stables pour la majorité des BPs. Le groupe (H, S, T et C) est le moins conservé. Cette faible association est principalement due à la cystéine qui peut être remplacée par d'autres acides aminés au sein des structures d'hélices α et des structures de transition entre les hélices α et les feuillets β (BPs *n* à *b*). De même, le groupe (A, L et M) n'est maintenu que pour 11 BPs sur 16.

Finalement, ces observations nous ont permis de proposer un alphabet réduit de 13 lettres dans lequel uniquement les associations fortes sont conservées : G, P, (I, V), (F, Y, W), A, L, M, E, (Q, R, K) (N, D), (H, S), T, C (voir Figure 37). L'autorisation d'associations moins stables peut mener à des alphabets réduits de 11, 9, 8 ou 5 lettres.

Le principal intérêt de la définition d'un alphabet réduit d'acides aminés est de donner la possibilité à l'expérimentateur de choisir de manière raisonnée des mutations qui n'auront qu'un impact limité au niveau de la structure. Or, comme nous l'avons déjà signalé, les associations entre acides aminés sont très sensibles à l'information utilisée. Nous pensons

donc que l'utilisation des structures locales peut fournir des résultats plus pertinents que les méthodes basées sur l'utilisation de la séquence seule. Des analyses d'expériences de mutations soulignent d'ailleurs l'intérêt de notre approche en permettant d'établir un lien entre des modifications fonctionnelles et de stabilité et des changements de groupe d'acides aminés. Il serait intéressant de réaliser des études plus systématiques pour vérifier et analyser ce lien.

3.2.5.3 Construction de mots structuraux et élaboration d'un dictionnaire de synonymes

Comme nous l'avons vu jusqu'ici, les BPs sont un outil très pertinent permettant des études approfondies de la relation structure-séquence. Une limite de cet alphabet est la taille relativement petite des fragments décrits. Cinq résidus permettent de décrire une petite hélice α ou un petit brin β (de Brevern et al. 2000) mais ne permettent pas de capturer des interactions et des corrélations à longue distance dans la séquence. Or, nous avons montré dans le paragraphe 3.2.2, l'existence de transitions préférentielles entre BPs (Tableau 6). Ces préférences impliquent un nombre limité d'arrangements privilégiés limité entre BPs. Plusieurs questions se posent alors :

- Les BPs restent-ils une description fiable pour la caractérisation des structures plus longues ?
- Quelles sont les règles logiques gouvernant l'association des BPs au sein des structures ?

Ainsi, en 2002, de Brevern et collaborateurs ont mené une étude sur des séries de 5 BPs (soit 9 résidus consécutifs) (de Brevern et al. 2002). Les BPs étant définis comme les *lettres* d'un alphabet structural, nous parlerons donc de *mots* pour nommer les séries de BPs. Cette analyse a permis de mettre en évidence les 72 mots structuraux les plus fréquents. Ces mots décrivent en moyenne plus de 92% des résidus des structures protéiques (voir Figure 38).

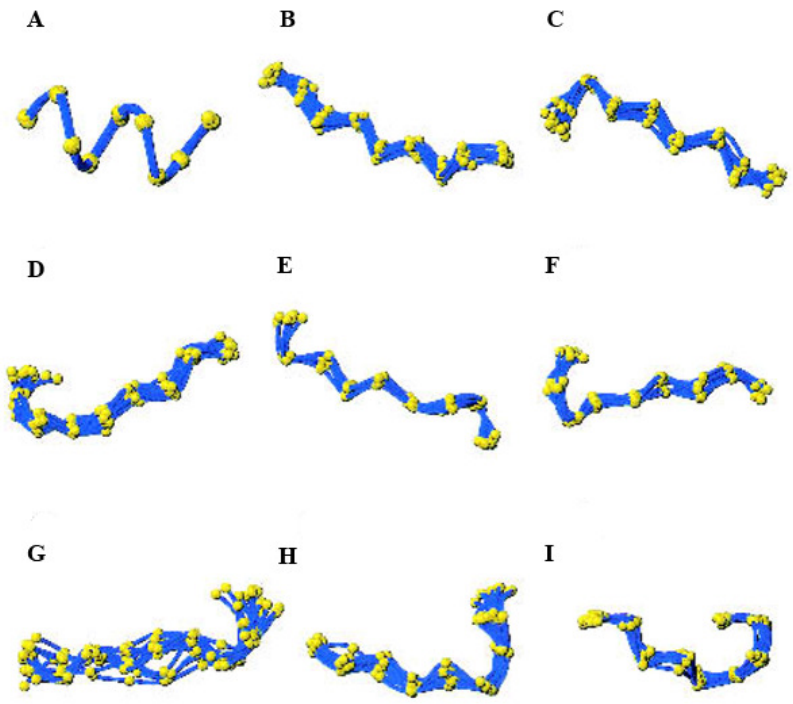

Figure 38. Exemples de fragments protéiques associés à neufs mots structuraux.

A – *mmmmm*, B – *ddddd*, C – *bccdd*, D – *pacdd*, E – *acddf*, F – *iacdd*, G – *cddfb*, H – *ddfbf*, I – *dfbfk*. La fréquence de ces mots est respectivement de (a) 17,27%, (b) 3,87%, (c) 0,37%, (d) 0,54%, (e) 0,20%, (f) 0,52%, (g) 0,24%, (h) 0,22%, (i) 0,22%. Figure extraite de (de Brevern et al. 2002).

Par ailleurs, un Cα RMSD de 0,7 Å est observé en moyenne entre deux fragments structuraux associés à un même mot. Cette valeur varie de 0,34 à 1,11 Å selon le mot considéré. Une bonne qualité d'approximation est donc obtenue pour ces longs fragments de 9 Cα de long. De plus, la plupart des mots identifiés se chevauchent : les quatre derniers BPs d'un mot peuvent être identiques aux quatre premiers BPs d'un autre. Ainsi, *mnopa* est chevauchant avec *nopac*, *nopab* et *nopaf*. Ce phénomène est notamment très fréquent au niveau des extrémités N et C terminales des structures étendues : *ccddd* peut transiter vers *cdddd*, *cdddf* ou *cddde*. En s'appuyant sur ces observations, un réseau orienté a été construit dans le but de caractériser les règles de transition entre BPs au sein des structures (voir Figure 39). Par exemple, il est possible de retrouver le mot *mnopa* suivi de *nopac*. Les 58 mots les plus fréquents ont été utilisés. Ainsi, le réseau caractérise 90% des résidus des structures protéiques (très similaires aux 92% avec 72 mots). Un point également très intéressant est que ce réseau décrit 80% des résidus ayant une conformation de type boucle (*i.e., coil*).

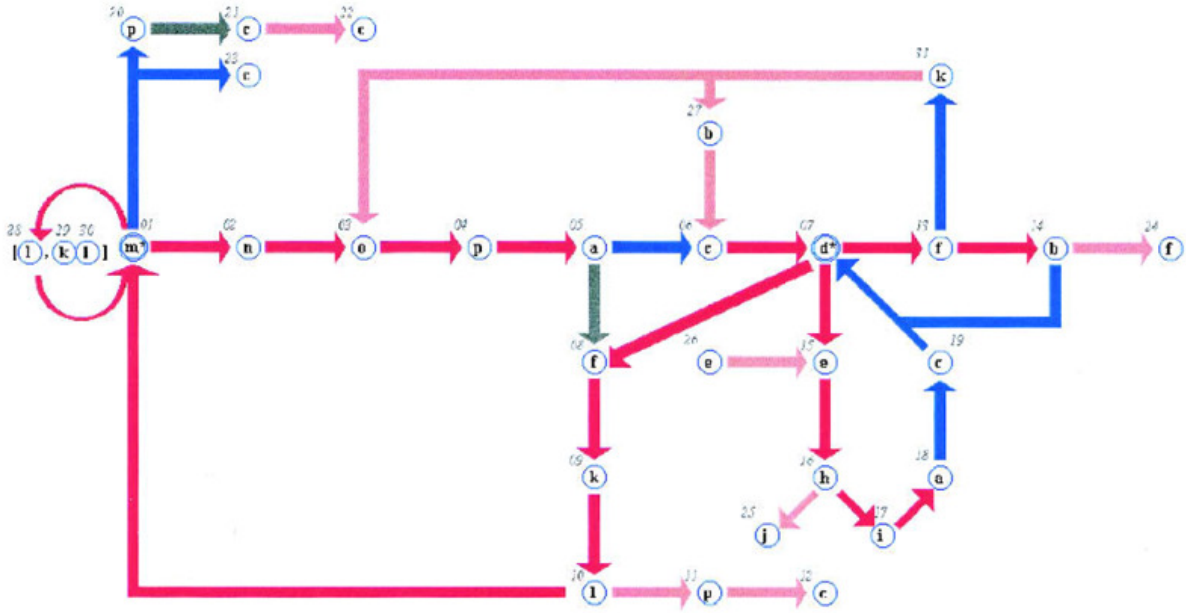

Figure 39. Réseau formé à partir des MSs.

Le réseau est constitué de 31 nœuds. Chaque nœud est numéroté et caractérisé par un BP. Les couleurs dépendent du nombre d'occurrences, avec (rouge) : > 10% de la banque de données, (bleu) : entre 10% et 6%, (gris) : entre 6% et 3,5%, et (rose) : < 3,5%. Les nœuds 01 et 07 correspondent aux BPs *m* et *d*. Leurs répétitions sont symbolisées par une étoile (*m** et *d**). Les nœuds 28 (BP *l*) et 29-30 (BPs *k* et *l*) sont inclus dans les successions *mlm* et *mklm*. Le réseau est discontinu, et par conséquent les nœuds 12, 22, et 23 (BP *c*), le nœud 25 (BP *j*) et le nœud 24 (BP *f*) sont des terminaisons. Figure extraite de (de Brevern et al. 2002).

Ces propriétés ont été utilisées dans le cadre du développement d'une méthode de prédiction de BPs. Les spécificités de séquence des 72 mots ont été analysées. A partir d'une séquence cible, l'adéquation séquence – structure avec les 72 mots structuraux sont calculées. Les meilleures prédictions sont directement identifiées grâce à un indice de confiance. Les propriétés de transitions sont ensuite utilisées pour étendre ces prédictions en amont et en aval. Un taux de prédiction (Q_{16}, proportion de BPs correctement prédits) égal à 38% a été obtenu, assurant un gain par rapport à la même stratégie appliquée aux BPs (Q_{16} = 34,4%).

Ainsi, ces études ont permis de montrer que les séries de BPs ont un sens structural et que les BPs sont fiables dans la caractérisation de longs fragments. De plus, des règles logiques d'assemblage des blocs au sein des protéines existent. Enfin, les spécificités de séquences des mots de 9 Cα de long sont suffisamment informatives pour permettre d'améliorer le taux de prédiction de BPs par une approche simple. Toutefois, dans ces analyses près de 10% des résidus ne sont pas pris en compte. De plus, les mots étudiés font 9 Cα de long. Or pour se diriger vers une meilleure compréhension de l'architecture des protéines et pour progresser vers la proposition de modèles protéiques dans le cadre de la prédiction *de novo*, il est

nécessaire de considérer des fragments de plus en plus longs. Comment étudier des mots plus longs alors que cet allongement va provoquer la multiplication des séries de BPs ? Une méthode probabiliste a été proposée : la méthode de la protéine Hybride (de Brevern and Hazout 2001). Cette méthode d'apprentissage propose de regrouper non plus des séries exactement identiques de BPs mais des séries similaires. Cette stratégie est à la base de la définition de la librairie des Prototypes de Structures Locales (PLSs). Cette dernière est présentée dans le paragraphe suivant. Les PLSs seront la matière première des développements méthodologiques qui constituent mon principal travail de thèse (voir sections 4 et 6).

3.3 La librairie de mots structuraux représentatifs ou «Prototypes de Structures Locales» (PSLs)

En 2006, dans le cadre de sa thèse au sein du laboratoire, Cristina Benros a développé une nouvelle librairie de longs fragments structuraux de 11 résidus de long (Benros 2005; Benros et al. 2006).

Dans un premier temps, ce paragraphe présentera la stratégie utilisée pour définir cette bibliothèque. Nous résumerons ensuite les caractéristiques structurales et les spécificités de séquences observées au sein des classes de fragments. Finalement, nous décrirons le développement d'une méthode de prédiction associée ainsi que les résultats obtenus.

Il est d'ores et déjà important de souligner que l'analyse et surtout la prédiction de fragments de structure de 11 résidus de long étaient des projets extrêmement ambitieux. Les plus longues structures locales analysées de manière systématique étaient alors seulement de 9 résidus (Schuchhardt et al. 1996).

Avec 11 résidus, il est possible de se rapprocher de la longueur moyenne des hélices (14 résidus ±5 (Kumar and Bansal 1998)) et d'englober les brins β (5 à 10 résidus (Branden and Tooze 1998)) ou encore les boucles longues (plus de 8 résidus). La prise en compte de longs fragments est un avantage majeur pour pouvoir capturer des interactions à longue distance et aller vers la construction de modèles protéiques structuraux globaux. Cependant, la classification et la prédiction deviennent plus difficiles du fait de l'augmentation de la variabilité structurale et de séquence.

3.3.1 Définition des PSLs

La bibliothèque développée par Benros et collaborateurs compte 120 classes structurales caractérisant des fragments de 11 résidus de long. Le nombre de classes et la longueur des

fragments ont été choisis de manière empirique. Ces deux paramètres sont liés. Le nombre de classes doit être :

(i) suffisant pour permettre une description précise des structures locales de 11 résidus de long.

(ii) mais limité pour permettre un peuplement satisfaisant des classes et la mise en évidence de spécificités de séquence exploitables pour la prédiction.

La classification des fragments est basée sur la méthode de la Protéine Hybride et sur l'analyse de séries de *L* BPs successifs. L'observation selon laquelle certains BPs partagent des structures locales proches a conduit au développement du concept de *mot structural « flou »*. Selon ce concept des séries de BPs non identiques mais similaires peuvent caractériser des structures locales proches. Ainsi, la Protéine Hybride (ou *Hybrid Protein Model* en anglais, HPM) est une méthode de classification non supervisée capable de créer *N* groupes de fragments caractérisés par des séries de *L* BPs similaires (de Brevern and Hazout 2001; Benros et al. 2003; de Brevern and Hazout 2003). *N* et *L* sont définis par l'utilisateur. Un mot structural n'est donc plus caractérisé par une série exacte de *L* BPs mais par un profil de distributions des 16 BPs en chacune des *L* positions. Le processus d'apprentissage de la HPM est proche de celui des cartes topologiques de Kohonen (Kohonen 1989; 1997) (voir aussi paragraphe 3.2.1). De plus, cette technique est spécialement dédiée à l'étude des structures protéiques. Ainsi, un avantage important est qu'elle force l'apprentissage d'une continuité entre les classes et définit donc des classes chevauchantes. En conséquence, la classe *i* regroupe des fragments:

(i) caractérisés par des séries de BPs similaires entre elles.

(ii) pour lesquels les *L-x* derniers BPs sont similaires aux *L-x* premiers BPs de la classe *i+x*. Par exemple, comme vu précédemment pour *L*=5, les mots de type *mnopa* en position *i* seront chevauchants avec les mots de type *nopac*, *nopab* ou *nopaf* en position *i*+1 (paragraphe 3.2.5.3).

(iii) pour lesquels les *L-x* premiers BPs sont similaires aux *L-x* derniers BPs de la classe *i-x*.

A la fin de l'apprentissage, le modèle créé est donc également représentatif de l'architecture des protéines puisque la continuité des fragments au sein des structures est retenue.

Concrètement, le principe général de la création de la bibliothèque des PLSs à donc été le suivant (voir Figure 40) :

- La longueur des séries de BPs *L* a été fixée à 7 (nombre de BPs nécessaires pour caractériser la conformation de fragments de 11 résidus de long).
- Le nombre de classes structurales *N* a été fixé à 120.
- Tous les fragments de 11 résidus de long d'une banque non-redondante de 675 structures protéiques ont été codés en termes de mots structuraux de 7 BPs.
- La classification des mots structuraux a été réalisée grâce à la HPM et a permis de définir 120 classes structurales chevauchantes.
- Pour chaque classe structurale, le fragment barycentrique en termes de Cα RMSD a été choisi comme représentant ou *Prototype de Structure Locale (PLS)*.
- Finalement, une dernière étape a consisté à s'assurer que tous les fragments de la banque étaient bien assignés au prototype qui leur était le plus proche (Cα RMSD). Une réassignation a été effectuée lorsque nécessaire.

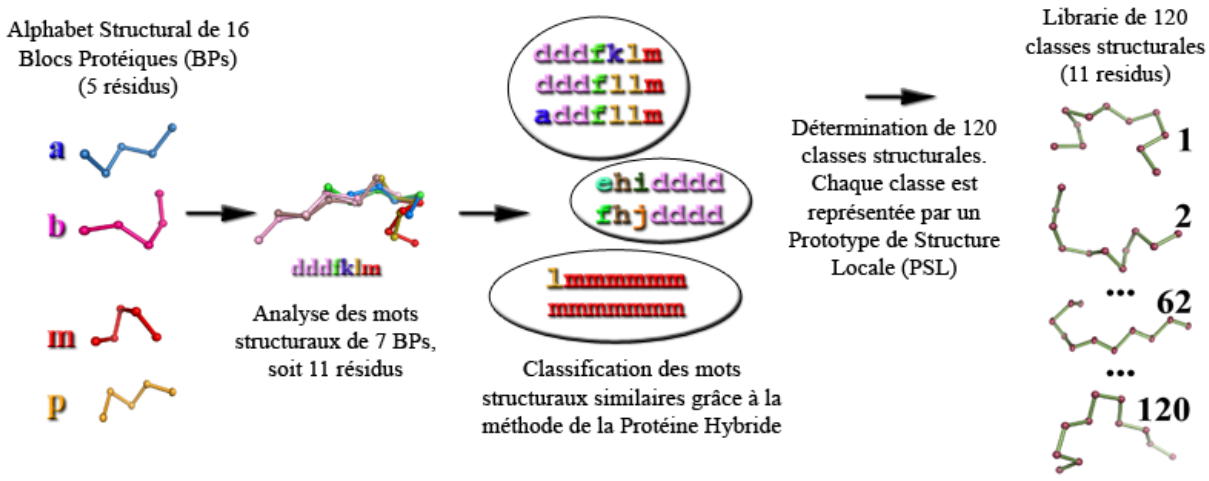

Figure 40. Principe général de la définition de la librairie.

Voir le texte pour les explications.

3.3.2 Caractéristiques structurales des PSLs

La Figure 41 présente la diversité structurale d'une partie des classes obtenues.

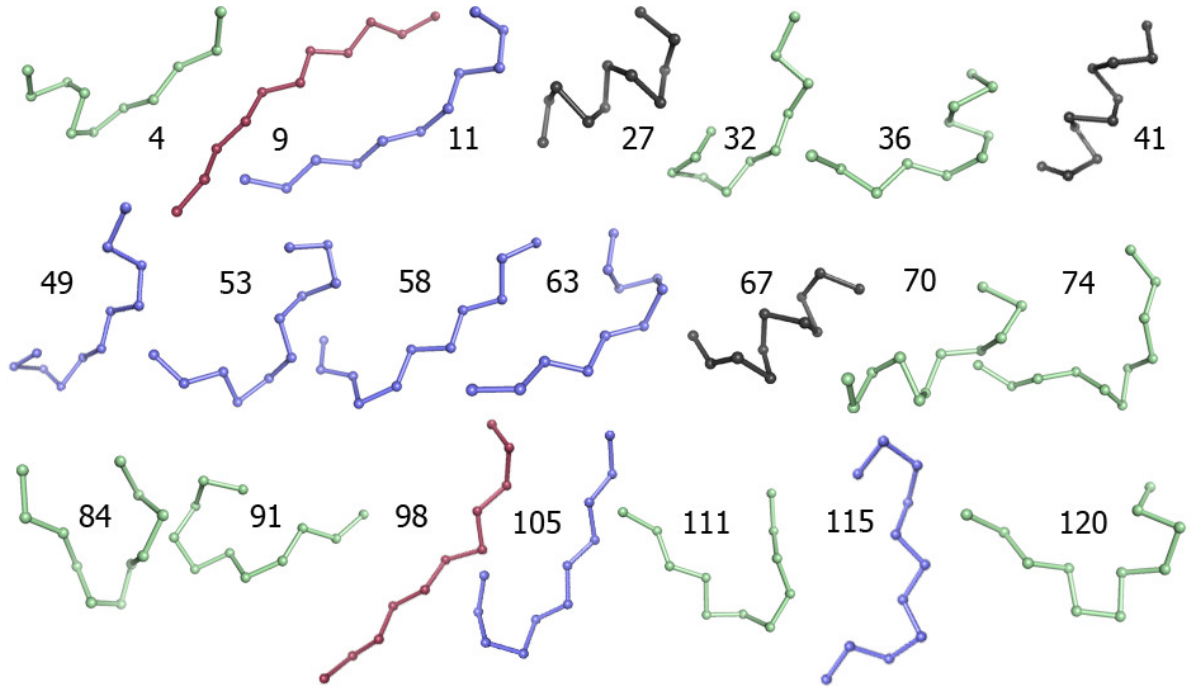

Figure 41. Exemples de PSLs.

Les conformations de 21 exemples de PSLs sont représentées. Chaque sphère correspond à un Cα. Pour la clarté de la représentation, les Cα sont reliés entre eux par des liaisons fictives. Les couleurs correspondent à quatre catégories de PSLs définies, pour analyse, en fonction des structures secondaires : les structures hélicoïdales (noir), les cœurs de structures étendues (rouge), les extrémités de structures étendues (bleu) et les structures de connexion (vert) (voir le texte pour les explications).

La bibliothèque des PSLs obtenue peut être caractérisée par un *profil structural*. Ce profil global correspond à la concaténation des classes structurales codées en termes de BPs (cf. Figure 42a). Il correspond au modèle de la Protéine Hybride. Il donne la probabilité de chacun des 16 BPs, de *a* à *p*, en chacune des 120 positions. La classe *i* est caractérisée par les *L*=7 distributions de PBs allant de *i*-3 à *i*+3. La classe *i*+1 est chevauchante et est définie de *i*-2 à *i*+4. Ces deux classes partagent donc 6 distributions en commun. Ainsi, parcourir la Protéine Hybride de la position 1 à la position 120 revient à parcourir de résidu en résidu une structure protéique modèle résumant l'architecture de toutes les protéines connues. Il est à noter que le modèle est circulaire fermé : la classe structurale 120 est suivie par la classe 1 au sein des structures protéiques.

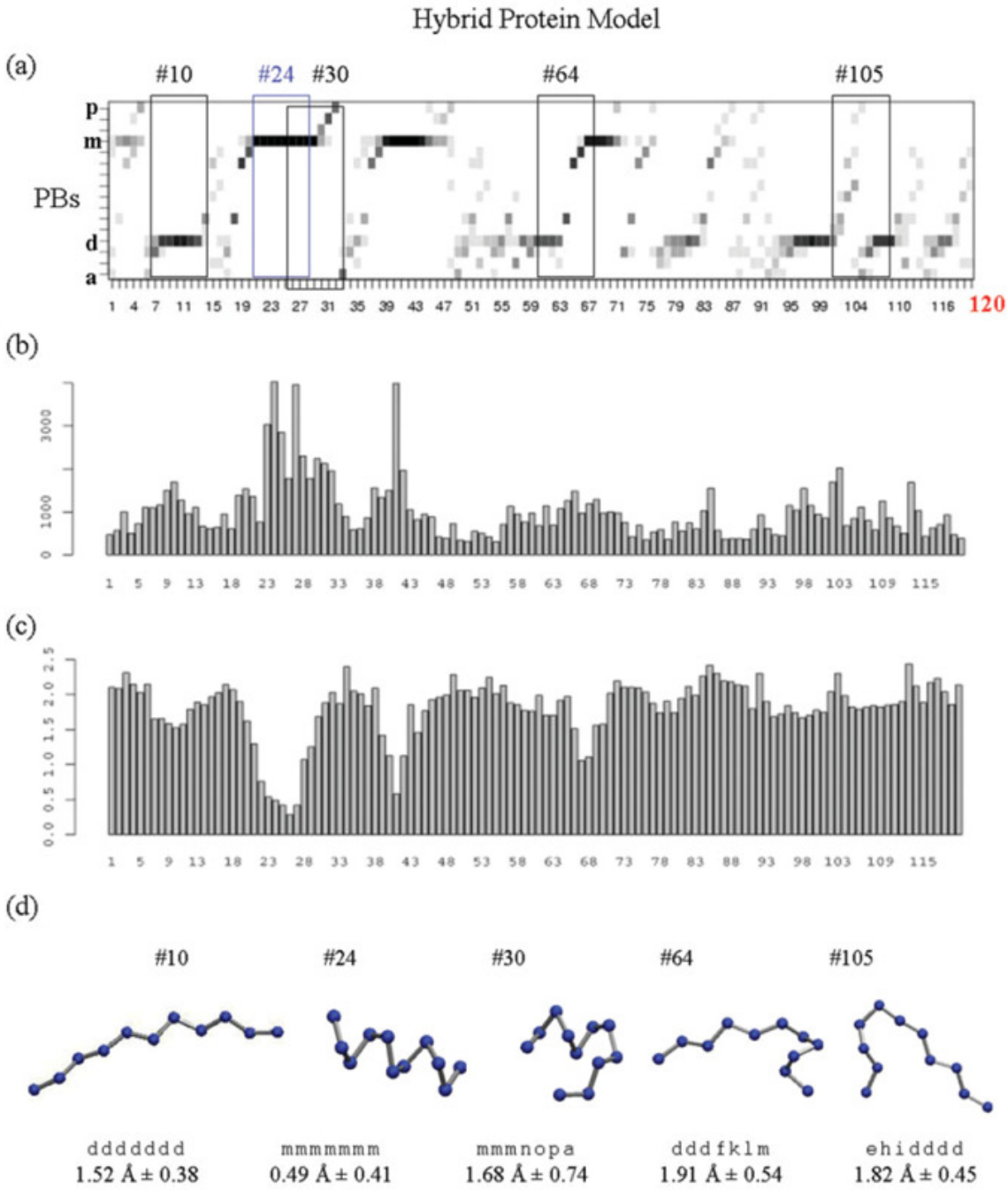

Figure 42. Propriétés de la bibliothèque de PSLs.

(a) La bibliothèque est caractérisée par un profil structural en BPs, *i.e.,* une série de distributions des 16 BPs. Une classe est définie par $L = 7$ distributions. Les classes successives sont chevauchantes. Les classes n°10, n°24, n°30, n°64 et n°105 sont représentées en exemples. (b) Répartition des fragments dans les différentes classes. (c) Valeurs des Cα RMSD moyens (Å). (d) Exemples détaillés de prototypes moyens représentés avec le logiciel VMD (Humphrey et al. 1996). Ce sont des fragments de 11 Cα, codés par des séries de 7 BPs. L'extrémité N-terminale se situe sur la gauche. Le Cα RMSD moyen de leur classe est donné ainsi que l'écart type. Figure extraite de (Benros et al. 2006).

L'analyse du profil structural obtenu révèle des caractéristiques intéressantes (cf. Figure 42a). Une forte spécificité de structure est observée pour chaque classe le long de la Protéine Hybride. En effet, seuls quelques BPs caractérisent chaque position. Par ailleurs de longues séries de BPs identiques sont observées. Ces séries impliquent principalement des BPs *m* et *d* et correspondent donc à des structures répétitives, hélices α et brins β respectivement (voir paragraphe 2.2.2) (Figure 42d, PSLs 24 et 10). Trois séries d'hélices et six séries de brins sont observables. Elles diffèrent par leur position dans l'architecture de la Protéine Hybride et donc par leur environnement local. Par exemple, la région hélicoïdale 21-29 est suivie par une seconde région hélicoïdale 39-45. A l'inverse, cette dernière semble préférentiellement suivie par une région étendue (50-64). Ces séries diffèrent également par leur longueur et par leur spécificité structurale. Par exemple, la série (76-83) est riche en BPs *d*, néanmoins, elle fait aussi intervenir les BPs *b* et *c* également associés à des structures étendues mais légèrement plus compactes (Tableau 6). Par ailleurs, entre les structures répétitives, certaines structures de connexions sont très bien définies comme les classes 30 ou 64 majoritairement définies par *mmmnopa* et *dddfklm* (cf. Figure 42a et d).

De plus, la variabilité *intra*-classe a été étudiée en superposant au mieux tous les fragments d'une classe sur leur PSL représentatif et en calculant une distance géométrique : le Cα RMSD. Le Cα RMSD moyen obtenu est de 1,61 Å (σ=0,77). Cette valeur indique une faible variabilité géométrique au sein des classes. En effet, l'approximation structurale obtenue en superposant deux fragments choisis au hasard dans la banque de données est bien moins précise pour des fragments de 11 résidus de long : 4,5 Å (σ=1,1). Le Cα RMSD varie en fonction des classes de 0,28 à 2,44 Å (Figure 42c). La classe hélicoïdale 26 bénéficie de la plus faible variabilité tandis que la classe 113, correspondant à une connexion entre 2 structures étendues, est associée à la plus forte.

Par ailleurs, la variabilité *inter*-classe a été analysée en superposant au mieux les 120 PSLs et en calculant le Cα RMSD entre chaque paire (cf. Figure 43). Cette étude montre que la redondance structurale est très faible au sein de la bibliothèque. La plupart des PSLs sont distants de plus 2,5 Å. La moyenne des Cα RMSD est de 3,9 Å (σ=1,0). Seules 1,8% des paires de PSLs présentent une distance géométrique inférieure à 2 Å. Comme attendu, ces paires correspondent principalement à des classes de structures répétitives.

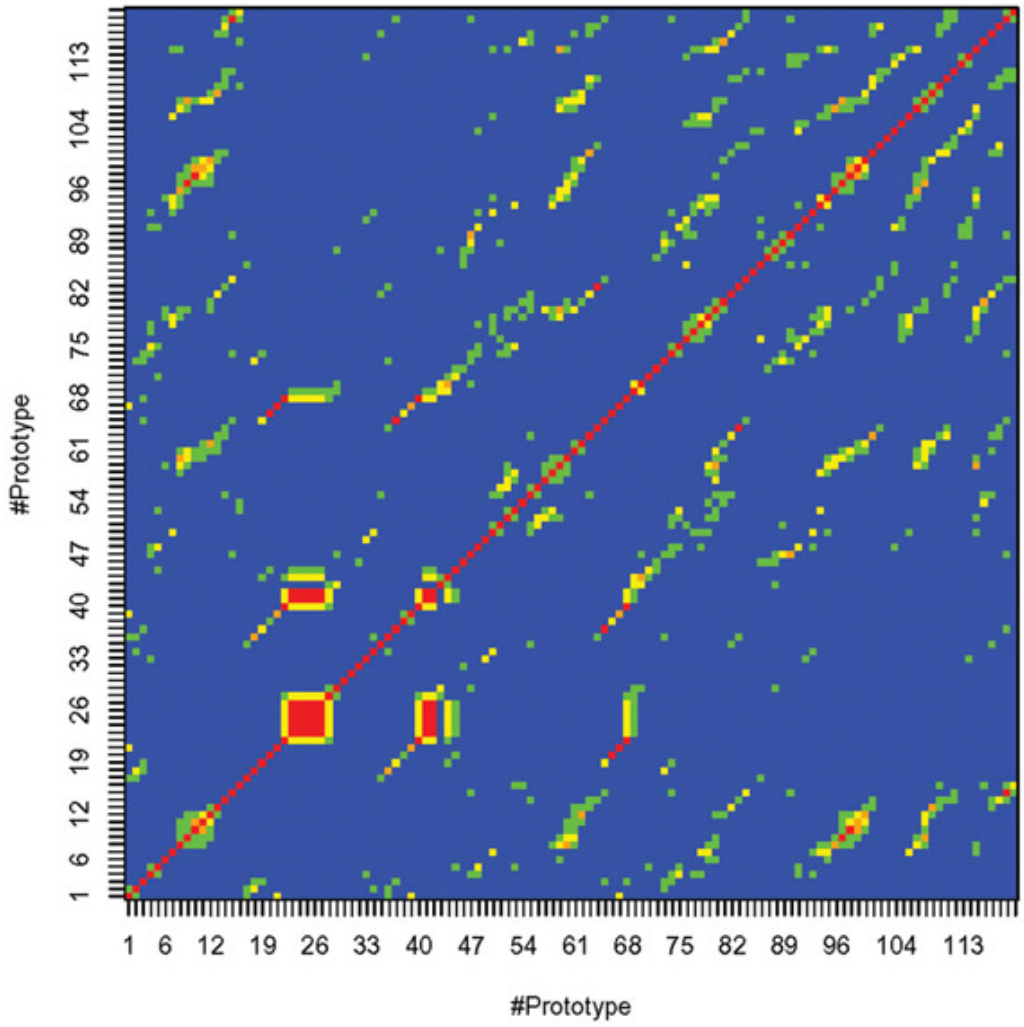

Figure 43. Comparaison structurale des 120 PSLs.

Les valeurs de *RMSD* entre Cα sont calculées après superpositions optimales des 120 prototypes deux à deux. Les couleurs sont affectées comme suit : (rouge) : Cα *RMSD* < 1 Å, (orange) : 1 Å ≤ $C\alpha$ *RMSD* < 1,5 Å, (jaune) : 1,5 Å ≤ $C\alpha$ *RMSD* < 2 Å, (vert) : 2 Å ≤ $C\alpha$ *RMSD* < 2,5 Å, et (bleu) : $C\alpha$ *RMSD* > 2,5 Å. Figure extraite de (Benros et al. 2006).

Enfin, la répartition des fragments par classe est présentée en Figure 42b. Le nombre de fragments associé à chaque classe varie de 308 à 4028. Chaque classe possède donc un effectif raisonnable. La première région hélicoïdale est la plus peuplée : la classe la plus peuplée est la n°24. A l'inverse, la moins peuplée est la 55, localisée dans une région étendue. Les 120 PSLs sont nécessaires pour donner une approximation correcte de toutes les structures locales connues. Cependant, dans un but d'analyse, Benros et collaborateurs ont défini quatre catégories de PSLs par classification hiérarchique. Globalement, ces catégories ont regroupé les classes structurales en fonction des structures secondaires qu'ils contiennent. Les quatre catégories sont : les structures hélicoïdales, les cœurs de structures étendues, les

extrémités de structures étendues et les structures de connexion (cf. Figure 41). Elles contiennent respectivement 16, 13, 40 et 51 PSLs.

En conclusion, la librairie des PSLs permet une bonne approximation structurale des fragments de 11 résidus de long observés dans les structures protéiques connues. Le nombre important de prototypes structuraux permet une description fine des structures locales et notamment des extrémités de structures répétitives. Il est également important de noter que tous les fragments protéiques on été pris en compte contrairement à l'étude menée précédemment (de Brevern et al. 2002). Enfin, les classes sont suffisamment peuplées pour permettre une étude de la relation séquence-structure locale. Le paragraphe suivant expose les résultats de ces analyses réalisées par Benros et collaborateurs (Benros et al. 2006).

3.3.3 Spécificité de séquence des PSLs

Après la construction de la bibliothèque de PSLs, Benros et collaborateurs ont analysé les spécificités de séquences associées à chaque classe structurale. Comme dans les études précédentes, les auteurs ont pris en compte des fenêtres de séquence étendues à chaque extrémité des fragments (paragraphe 3.2.3). Ainsi, des séquences de 21 résidus ont été considérées. Elles seront notées de -10 à +10 et centrées en 0. Comme dans les études précédentes également des matrices d'occurrences en acides aminés ont été calculées (paragraphe 3.2.3). Elles ont ensuite été normalisées comme suit :

$$i^s(AA_k^j) = \ln\left[\frac{p_{obs}^s(AA_k^j)}{p_{th}(AA_k)}\right]$$

$p_{obs}^s(AA_k^j)$ est la probabilité d'observer l'acide aminé AA_k en position j de la fenêtre de séquence pour la classe de PSL s d'intérêt. $p_{th}(AA_k)$ est la probabilité d'observer l'acide aminé AA_k. $i^s(AA_k^j)$ est donc le logarithme d'une fréquence relative. Il est positif si l'acide aminés considéré est sur-représenté et négatif s'il est sous-représenté. De plus, pour évaluer l'informativité contenue en chaque position des séquences, la mesure du *KLd* a été calculée (paragraphe 3.2.3). Plus cette mesure est élevée, plus la position est informative. La significativité statistique du *KLd* a été évaluée par rapport à un risque de première espèce α de 10^{-3}.

Ainsi, le nombre de positions significativement informatives pour chaque classe structurale est en moyenne de 12,6 et peut varier de 4 (classes n° 50 à 53) à 20 (classe n° 27). Ces

positions sont majoritairement localisées de -5 à +5, *i.e.* dans la région centrale. En moyenne, 9,4 des 11 positions centrales sont significativement informatives. La Figure 44 illustre les PSLs pour les classes 67 et 15 (a), leur matrice d'occurrences normalisée (b) et la distribution du *KLd* associée (c). La classe 67 caractérise l'extrémité N-terminale d'une hélice et bénéficie d'une faible variabilité structurale (Cα RMSD 1,06 Å). Le *KLd* montre de plus 14 positions significatives en termes de séquence. Les positions les plus significatives sont contenues dans la région centrale. Par exemple, en position -3, les acides aminés (P, G S, T, D) sont sur-représentés tandis que les acides aminés hydrophobiques sont sous-représentés. La classe structural 15 est structuralement légèrement moins déterminée que la classe 67 (Cα RMSD 1,96 Å). Néanmoins, elle bénéficie de 8 positions significativement informatives en terme de séquence. Les acides aminés I et V sont par exemple sur-représentés en positions -4 à -2.

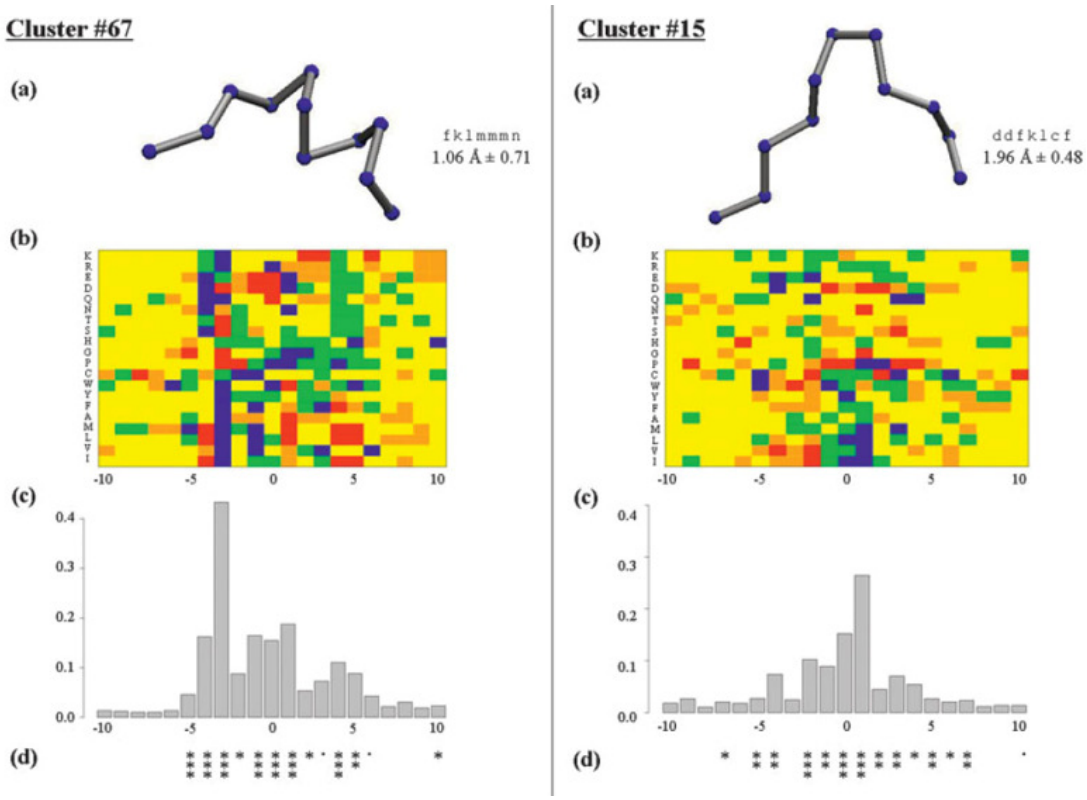

Figure 44. Analyse de la relation séquence-structure locale.

(a) Le prototype moyen est représenté, et sont donnés : son codage en BPs, son Cα *RMSD* moyen d'approximation structurale (*intra*-classe), et l'écart type correspondant. (b) La matrice d'occurrences en acides aminés normalisée est représentée. Les couleurs sont affectées en fonction de la distribution des valeurs normalisées de l'ensemble des 120 matrices d'occurrences comme suit : 5% des valeurs les plus élevées (rouge), 15% suivantes (orange), 5% des valeurs les plus faibles (bleu), 15% suivantes (vert), autrement (jaune). (c) Distribution des valeurs de *KLd*. Le seuil de significativité est calculé pour correspondre à une risque a de première espèce de 10^{-3}. Au dessus de cette valeur, les positions sont informatives. Figure extraite de (Benros et al. 2006).

3.3.4 Prédiction des PSLs à partir de la séquence

Ainsi, de fortes relations séquence-structure ont été mise en évidence au sein des classes de structures locales. Ces relations ont ensuite été exploitées pour mettre en place une stratégie de prédiction des fragments de structure de 11 résidus de long (Benros et al. 2006).

3.3.4.1 Stratégie de prédiction

Le principe de la méthode de prédiction développée par Benros et collaborateurs est présenté en Figure 45. Il repose sur un système d'*experts*. Pour chaque classe structurale *s* représentée par son PSL, un *expert* apprend à séparer de façon optimale les fragments de séquences associés à *s* (exemples positifs) des fragments associés aux autres classes (exemples négatifs) (cf. Figure 45e). Pendant la prédiction, chaque expert va calculer un score de compatibilité entre sa classe (la classe pour laquelle il a été entrainé) et la fenêtre de séquence cible (Figure 45f). Les 120 scores sont ensuite ordonnés. Finalement, un jury sélectionne les classes ayant obtenu le meilleur score de compatibilité et propose les PSLs associés comme candidats structuraux (cf. Figure 45g). De plus, un indice de confiance est calculé et permet d'évaluer directement la prédiction.

La banque de données utilisée pour l'apprentissage et la validation de la méthode de prédiction est une banque réactualisée de 1041 structures protéiques non redondantes. Trois partitions ont été effectuées : l'échantillon 1 est utilisé pour apprendre les spécificités de séquence de chaque classe (521 protéines, 125 074 fragments), l'échantillon 2 est utilisé pour un calibrage des experts (261 protéines, 62 194 fragments), enfin, l'échantillon 3 permet une validation de la méthode (259 protéines, 64 229 fragments). Ces différentes étapes sont présentées dans la suite du texte.

Mise en place du système d'expert

Les 120 experts sont définis par régression logistique. Une fonction logistique va permettre de calculer la probabilité $p(W/s)$ qu'une fenêtre de séquence *W* appartienne à une classe *s* de fragments structuraux. Elle est définie comme suit :

$$p(W/s) = \frac{1}{1+\exp[-\phi_s(W)]} \qquad \text{avec} \qquad \phi_s(W) = \omega_0 + \sum_{j=-m}^{j=+m} \omega_j . i_s(AA_k^j)$$

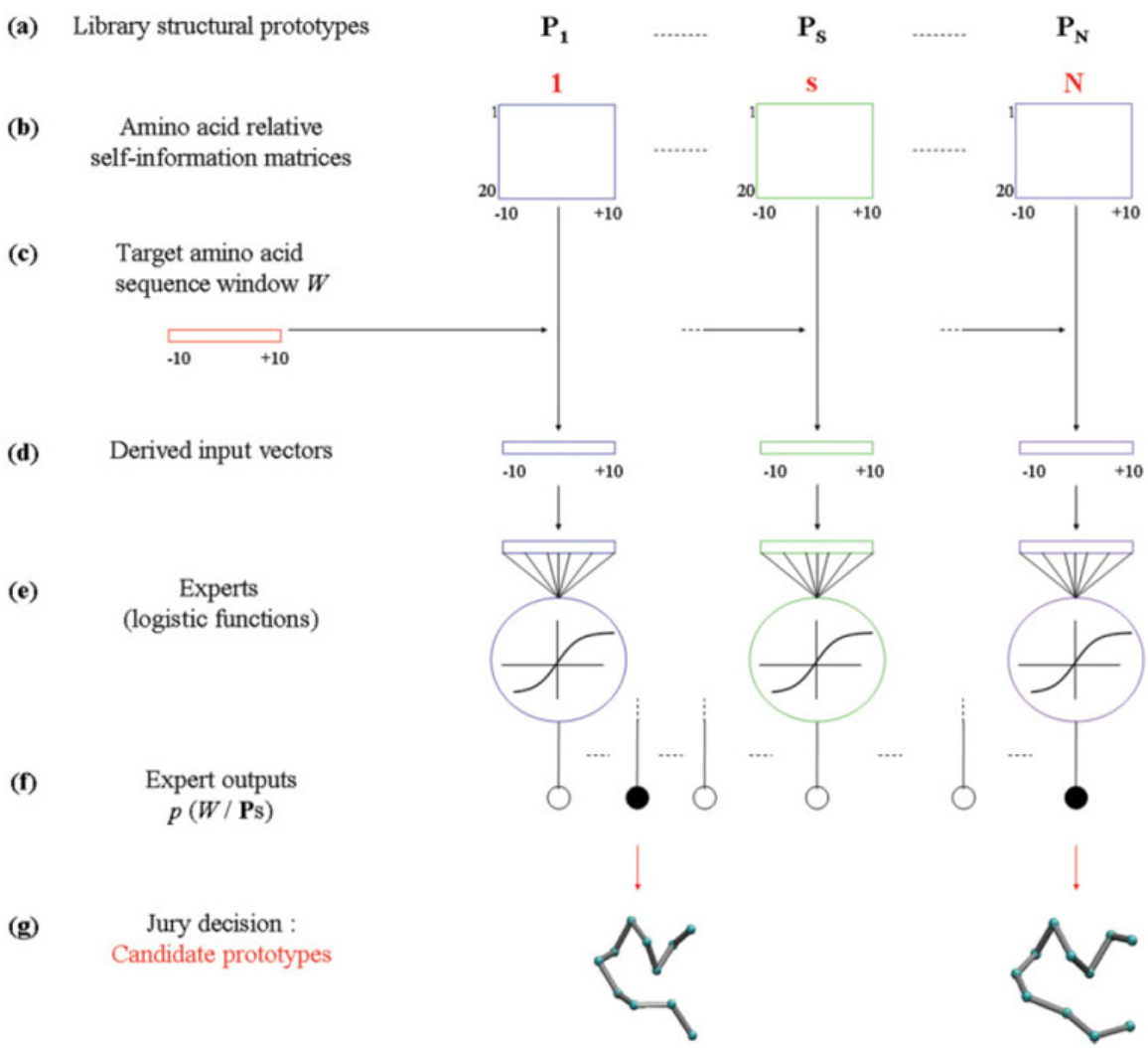

Figure 45. Stratégie de prédiction des PSLs.

(a) Chaque classe structurale *s* de la bibliothèque, *s* variant de 1 à *N*=120, est représentée par un PSL. (b) Pour chaque classe, une matrice d'occurrences en acides aminés est calculée puis normalisée. Elle est de dimension (20x21), correspondant aux 20 acides aminés et à des fenêtres de séquences de 21 résidus de long (numérotées de -10 à +10). (c) Pour chaque fenêtre de séquence cible *W*, d'une longueur de 21 résidus et de structure inconnue, (d) un vecteur d'informations est dérivé pour chacune des *N* classes de la bibliothèque. (e) Les vecteurs résultants constituent les données d'entrée des *N* experts. Chaque expert est caractérisé par une fonction logistique. (f) La sortie de chaque expert correspond à la probabilité que la fenêtre de séquence cible soit compatible avec le prototype de la classe auquel il est associé, à savoir $p(W / \mathbf{P}_s)$. Ces sorties sont analysées par un jury qui, basé sur une règle de décision, va proposer ou non le prototype en tant que candidat structural (g). Dans l'exemple, deux prototypes candidats sont proposés. Figure extraite de (Benros et al. 2006).

Au sein du modèle (ou fonction) associé à la classe *s*, les séquences seront décrites par un vecteur de $i_s(AA_k^j)$, ce sont les *variables explicatives* (voir paragraphe 3.3.3) (Figure 45b,c,d). Ces valeurs sont extraites de la matrice d'occurrences normalisée associée à *s*. Elles caractérisent en chaque position *j* la probabilité d'observer l'acide aminé *k* présent dans la séquence à analyser. Ainsi, en fonction de la classe *s* considérée, une même séquence n'aura pas le même vecteur de variables explicatives, chaque classe ayant une matrice d'occurrence en acides aminés spécifique (cf. Figure 45a). ω_0 et ω_j sont les coefficients du modèle. Ils devront être optimisés pour chaque position *j* de la fenêtre de séquence (indexée de $+m$ à $-m$). Les poids présentent l'intérêt majeur de permettre l'évaluation de la contribution des acides aminés localisés aux différentes positions *j* dans le pouvoir discriminatif de l'expert. Il est de plus important de noter que les fonctions logistiques permettent de calculer des valeurs de $p(W/s)$ comprises entre 0 et 1.

Ainsi, le calcul des matrices d'occurrences a été réactualisé en utilisant l'échantillon 1 et l'optimisation des poids ω par régressions logistiques a été réalisée sur l'échantillon 2 de calibrage. Dans ce but, à partir de cet échantillon 2, deux sous-échantillons ont été définis pour chaque classe *s*. Le premier sous-échantillon, dit *positif*, est constitué des fragments structuraux appartenant à la classe *s*. Le second sous-échantillon, dit *négatif*, est constitué du même nombre de fragments tirés aléatoirement dans les autres classes structurales et présentant au minimum une différence géométrique (Cα RMSD) de plus de 1,5 Å avec le PSL de la classe *s*. A chaque fragment structural correspond une fenêtre de séquence *W* qui sera codée en termes de $i_s(AA_k^j)$. A la fin de la régression, les poids ω associés à la classe *s* permettent de discriminer de manière optimale les fragments des échantillons positifs et négatifs.

Jury et règle de décision

Une fois les experts entrainés, la prédiction peut-être réalisée pour toute nouvelle fenêtre de séquence de 21 résidus de long. Pour chaque nouvelle séquence, chacun des 120 experts va calculer une probabilité $p(W/s)$ mesurant la compatibilité séquence-classe structurale.

Un jury va ensuite sélectionner les meilleurs PSLs candidats à partir des probabilités $p(W/s)$ ordonnées en ordre décroissant. La règle de décision utilisée pour sélectionner les candidats structuraux comporte deux volets : *(i)* la valeur $p(W/s)$ associée aux candidats doit être

supérieure à une valeur seuil de 0,8, *(ii)* le nombre de candidats sélectionnés n'excède jamais 5.

Evaluation des candidats structuraux

Cette stratégie de prédiction a été testée sur l'échantillon 3 de validation. Deux types d'évaluation ont été réalisés. Tout d'abord, de façon classique, une prédiction est définie comme correcte si le PSL assigné à partir de la structure est retrouvé parmi les candidats (Q_{120}). La seconde méthode d'évaluation est basée sur un critère géométrique (Cα RMSD) : une prédiction est jugée correcte si l'un au moins des PSLs candidats permet une approximation de la structure locale réelle meilleure qu'un seuil donné. Trois seuils ont été évalués : 1,5, 2 et 2,5 Å. Il est important de noter que le dernier seuil de 2,5 Å reste très strict. En effet, comme déjà exposé au paragraphe 3.3.2, l'approximation structurale obtenue en superposant deux fragments choisis au hasard dans la banque de données est de 4,5 Å. De même, la probabilité de trouver deux fragments présentant une similarité de moins de 2,5 Å est de 10^{-2}.

De plus, les auteurs ont également évalué l'efficacité de leur prédiction face à une prédiction dite *aléatoire*. Cette prédiction est réalisée pour chaque fragment de l'échantillon 3 en tirant au hasard le même nombre de candidats structuraux parmi les PSLs de la bibliothèque.

Construction d'un indice de confiance (IC)

Enfin, un indice de confiance a été défini pour évaluer directement les candidats structuraux proposés. Cet indice a été construit dans l'intention d'évaluer la probabilité qu'une prédiction soit correcte (selon le critère géométrique et pour un seuil de 2,5 Å). Ainsi, une nouvelle fonction logistique a été optimisée pour reconnaître les prédictions correctes des prédictions incorrectes à partir des données suivantes :

- Les types d'acides aminés présents dans la séquence cible (G, P, Hydrophobes, Polaires).
- Les PSLs choisis lors de la prédiction.
- La distribution des $p(W/s)$ obtenues.

A l'issue de l'apprentissage, la fonction logistique obtenue est capable de calculer la probabilité que toute nouvelle prédiction soit correcte. Pour définir l'IC, six classes de probabilités ont été délimitées. Elles correspondent à six niveaux de confiance : un IC de 1 indique un niveau de confiance faible, un IC de 6 est associé à un niveau de confiance élevé.

3.3.4.2 Résultats

Le taux de prédictions correctes (ou Q_{120}) obtenu en considérant uniquement la classe structurale assignée est égal à 35,0 %. Ce taux est déjà tout à fait satisfaisant étant donnée le nombre élevée de classes et la taille importante des fragments. De plus, il est largement supérieur au taux de 5,1 % obtenu avec une prédiction *aléatoire*.

Selon l'évaluation basée sur un critère géométrique, les taux de prédiction obtenus sont de 22,2 %, 35,1 % et 51,2 % pour les seuils d'approximation maximale autorisés de 1,5, 2 et 2,5 Å respectivement (paragraphe 3.3.4.1, Evaluation des candidats structuraux). Ces résultats correspondent à un gain de 17,0 %, 24,3% et 29,3 % respectivement par rapport à une prédiction *aléatoire*. Le Tableau 7 résume les contributions des différentes catégories de PSL dans ces résultats de prédiction. Les structures hélicoïdales (*Helical* en anglais dans le tableau) sont les mieux prédites quelque soit le seuil considéré. Des taux de 53,1 et 70 % sont obtenus en considérant des seuils de 1,5 et 2,5 Å respectivement. Le taux de prédiction est également élevé pour les cœurs de structures étendues (*extended core*) : un taux de 52,6 % est obtenu avec un seuil de 2,5 Å. Les PSLs associés aux deux dernières catégories étant ceux qui présentent la plus forte variabilité intra-classe, le seuil de 2,5 Å est le plus réaliste. Ainsi, des taux tout à fait satisfaisants de 43,3 et 42,4 % de prédiction correctes sont obtenus pour les extrémités de structures étendues (*extended edge*) et pour les structures de connexion (*connecting structures*) respectivement. Enfin, l'analyse des résultats de prédiction par classe structurale a permis de montrer que chacune participait aux taux obtenus. Toutes les classes sont prédites.

Tableau 7. Analyse de la prédiction par catégorie de PSL.

Prototype category	No. of prototypes	Proportion of fragments	Mean number of candidates ± *sd*	Prediction rate (%)		
				1.5 Å	2 Å	2.5 Å
Helical	16	26.9	3.9±1.3	53.1	63.2	70.1
Extended core	13	10.6	4.2±1.2	12.0	31.7	52.6
Extended edges	40	23.8	4.3±1.2	7.1	21.3	43.3
Connecting structures	51	38.7	4.3±1.2	12.9	24.9	42.4
All the library	120	100.0	4.2±1.2	22.2	35.1	51.2

Tableau extrait de (Benros et al. 2006).

Finalement, l'évaluation de l'indice de confiance a permis de montrer sa pertinence pour prédire la confiance qu'il est possible d'accorder aux prédictions. L'IC est effectivement

fortement corrélé au taux de prédiction. Ainsi, un IC de 1 est associé à un taux de prédiction de 17,4% alors qu'un IC de 6 correspond à un taux de prédiction de 85,8 % (critère géométrique, seuil de 2,5 Å).

3.3.4.3 Comparaison avec d'autres méthodes

La comparaison d'une stratégie de prédiction des structures locales avec une autre est loin d'être triviale. En effet, les définitions des structures locales peuvent être très diverses en fonction des alphabets utilisés. Comme nous l'avons vu dans le paragraphe 3.1.1, des différences majeures peuvent exister notamment au niveau du nombre de fragments représentatifs sélectionnés ou de leur longueur. De plus, peu d'alphabets structuraux sont mis à disposition de la communauté scientifique et encore moins sont associés à une méthode de prédiction. Ainsi, pour une analyse plus simple et plus pertinente, Benros et collaborateurs ont choisi de comparer les résultats de leur stratégie de prédiction à des méthodes de prédiction des angles dièdres Φ-Ψ (Benros 2005; Benros et al. 2006).

En 2000, Bystroff et collaborateurs ont développé un modèle de Markov caché (HMMSTR) exploitant les propriétés de chevauchement des I-sites pour la prédiction des structures locales (voir paragraphe 3.1.2.1). Cette méthode a été décrite et évaluée en termes de prédiction des angles de torsion du squelette polypeptidique. Un alphabet de 11 états conformationnels a été considéré (Bystroff et al. 2000). Ces états caractérisent 11 régions d'angles de torsions Φ-Ψ, définies sur une carte de Ramachandran (voir Figure 5 of Bystroff *et al.* (Bystroff et al. 2000), voir également la Figure 6 dans ce document). De même, Yang et Wang (Yang and Wang 2003) puis Kuang et collaborateurs (Kuang et al. 2004b) ont défini des méthodes dédiées à la prédiction de la conformation locale des résidus au sein de fragments de 9 résidus de long. Pour caractériser cette conformation locale, ils définissent 4 états : A, B, G et E. Dans un but de comparaison, Kuang a regroupé les 11 états de Bystroff en 4 états correspondant aux 4 états qu'ils avaient définis précédemment (Kuang et al. 2004b). De même, Benros a extrait les angles Φ-Ψ définissant chaque PSL et a codé ces derniers en fonction des 4 états conformationnels A, B, G et E. Dans ce cadre, la méthode de prédiction des PSLs a permis d'obtenir un taux de prédiction allant de 64 à 76% de prédictions correctes. Ce résultat, obtenu à partir de la seule séquence en acides aminés, est comparable aux 75% obtenus par Yang et Wang ainsi qu'aux 77% obtenus par Kuang et collaborateurs. Ces derniers bénéficient pourtant de la combinaison de leur méthode avec les résultats de prédiction des structures secondaires de PSIPRED (Jones 1999). Les résultats de la méthode de prédiction

des PSLs sont également comparables aux 74 % de prédictions correctes obtenues avec HMMSTR.

3.4 Conclusion

Au cours de ce chapitre, nous avons montré que les alphabets structuraux sont des outils puissants de description et de prédictions des structures locales protéiques. Les BPs développés en 2000 par de Brevern et collaborateurs ont déjà été utilisés pour de nombreuses applications par différents laboratoires. Les PSLs sont plus récents et représentent un défi ambitieux du fait de leur longueur très importante. Cependant, ils sont également un pas en avant vers la caractérisation et la prédiction de modèles protéiques globaux. En effet, la prédiction des PSLs développée par Benros et collaborateurs donne des résultats tout à fait prometteurs. Preuve que la recherche progresse et que nous parvenons à prédire des fragments de plus en plus longs.

La méthode mise en place pour la prédiction des PSLs est tout à fait originale et innovante. Cinq candidats structuraux au maximum sont prédits au lieu d'un seul classiquement. Cette stratégie permet de prendre en compte une certaine plasticité structurale des séquences dont la conformation peut être influencée par des interactions à plus longue distance. Elle permet également de réduire significativement la combinatoire (5 candidats au lieu de 120). Elle présente donc un réel intérêt pour les méthodes de prédiction de modèles protéiques globaux par assemblage de fragments.

Comme nous l'avons vu cette prédiction des PSLs est réalisée à partir de la séquence seule et repose sur des fonctions logistiques. Or, l'utilisation de données évolutionnaires et de méthodes d'apprentissage plus sophistiquées a déjà permis des améliorations importantes de méthodes de prédiction. La prédiction de structures secondaires ou encore des BPs notamment ont déjà largement bénéficié de ce type de stratégies (voir paragraphes 2.2.7 et 3.2.4).

Ainsi, mon premier travail de thèse, initié au cours de mon master 2, a été de répondre à la problématique suivante : est-il possible d'améliorer encore la prédiction des PSLs ? Le chapitre suivant décrit donc la stratégie utilisée pour optimiser la prédiction ainsi que l'évaluation des résultats les plus récents.

4. Nouvelle Stratégie de Prédiction des Structures Locales Protéiques (Article 9)

Dans ce chapitre, je présenterai tout d'abord le développement d'une nouvelle stratégie de prédiction plus performante pour les PSLs. Puis, dans un deuxième temps, je décrirais le développement d'un nouvel indice de confiance permettant d'évaluer directement la pertinence de la prédiction en chaque position de la séquence.

4.1 Une nouvelle stratégie pour améliorer la prédiction des structures locales

4.1.1 Objectifs

La méthode de prédiction originale des structures locales de 11 résidus de long a permis d'obtenir un taux de prédiction global de 51,2 % (Benros et al. 2006). Cette stratégie s'appuyait sur l'utilisation de la séquence seule et de fonctions logistiques (voir paragraphe 3.3.4).

Dans le cadre de ma thèse, j'ai mis en place une nouvelle stratégie de prédiction pour améliorer la prédiction des PSLs. Cette méthode bénéficie de la prise en compte de l'information de séquence contenue dans des protéines homologues à la séquence cible. Par ailleurs, une méthode d'apprentissage performante a été choisie : les Machines à Vecteurs Supports (*Support Vector Machines* en anglais, SVMs). Les SVMs ont déjà prouvé à plusieurs reprises leur efficacité dans le cadre de la prédiction de structures locales (Ward et al. 2003; Eddy 2004; Sander et al. 2006).

La mise en place de cette amélioration de la prédiction des PSLs a été l'un des axes principaux de mon travail de thèse. Je présenterai dans un premier temps le développement de la nouvelle méthode de prédiction. Puis, je décrirai les résultats obtenus. Une attention particulière a été portée sur l'obtention d'une prédiction équilibrée sans biais statistique. Enfin, les résultats de cette prédiction des structures locales seront discutés et comparés à d'autres stratégies de prédiction.

4.1.2 Méthodes

4.1.2.1 Stratégies mises en œuvre pour une amélioration de la prédiction

Deux types de stratégies ont été mises en œuvre pour tenter d'améliorer la prédiction des structures locales (Tableau 8) :

- L'utilisation d'experts définis par SVM en place de ceux définis par régression logistique.
- L'utilisation d'informations évolutionnaires.

Tableau 8. Nouvelles stratégies pour la prédiction des structures locales.

	Séquence seule	Données Evolutionnaires
Régression Logistique	✓	✓
SVM	✓	✓

En vert, est présentée la stratégie initiale mise en place par Benros *et al* (paragraphe 3.3.4.1) (Benros et al. 2006). En rouge, les différentes stratégies développées durant ma thèse.

Ainsi, dans le but d'évaluer précisément l'impact de l'utilisation des SVMs et/ou d'information évolutionnaires, le principe général de prédiction mis en place précédemment a été conservé. Chaque classe structurale *s* est associée à un expert entrainé à séparer les fragments de séquence associés à *s* des autres fragments. Lors de la prédiction pour une séquence cible donnée, chaque expert calcule un score de compatibilité avec sa classe. Un jury sélectionne ensuite les 5 meilleurs candidats.
Il est important de noter que le nombre de candidats structuraux a été fixé à 5 dans cette étude. Ce nombre fixe de candidats permet une comparaison plus aisée des différents schémas de prédiction.

4.1.2.2 Banque de structures protéiques

Pour permettre une comparaison directe, la banque de structures utilisées dans l'étude précédente a été conservée. Cette banque non redondante contient 1041 structures cristallographiques ayant une résolution meilleure que 2 Å (paragraphe 3.3.4.1). Elles partagent moins de 30% d'identité de séquence et diffèrent géométriquement de plus de 10 Å

(Cα RMSD) (cette sélection a été réalisée grâce au service web PDB-RPRDB (Noguchi et al. 2001)).

Chaque structure protéique est codée en termes de PSLs selon le critère du Cα RMSD minimal. Au total, 251 497 fragments sont obtenus. La partition de cette banque en trois échantillons distincts a été utilisée de la même façon que l'avaient fait Benros *et al.* (paragraphe 3.3.4.1):

- L'échantillon 1 comprend 521 protéines, soit 125 074 fragments.
- L'échantillon 2 comprend 261 protéines, soit 62 194 fragments. Il est dédié à l'apprentissage des experts.
- L'échantillon 3 comprend 259 protéines, soit 64 229 fragments. Il est dédié à l'évaluation des performances de la méthode de prédiction.

L'échantillon 1 était précédemment dédié à l'apprentissage de la relation séquence-structure locale, *i.e.,* au calcul des matrices d'occurrence. Nous verrons que notre nouvelle stratégie rend ce calcul inutile. L'échantillon 1 n'est donc plus utilisé ici. Il servira néanmoins lors de la mise en place de l'indice de confiance (paragraphe 4.2).

Comme précédemment, lors de l'entrainement des experts, l'échantillon 2 d'apprentissage sera redécoupé pour chaque classe structurale *s* (paragraphe 3.3.4.1). Ainsi, l'entrainement de l'expert associé à la classe *s*, est réalisé sur un échantillon composé pour moitié de fragments assignés à la classe *s* (exemples positifs) et de fragments choisis aléatoirement dans d'autres classes (exemples négatifs).

4.1.2.3 Construction du système d'experts

4.1.2.3.1 Enrichissement de la séquence en acides aminés par des données évolutionnaires.

Pour chaque protéine cible, le logiciel PSI-BLAST (Altschul et al. 1997) a été utilisé pour rechercher des séquences homologues dans la base de séquence non-redondante SWISSPROT (Boeckmann et al. 2003). Cet algorithme fonctionne de manière itérative. Il cherche tout d'abord des séquences similaires à la séquence cible. Puis, une Matrice de Score Position-Spécifique est calculée (*Position-Specific Scoring Matrices* en anglais ou PSSM). Elle contient des scores de sur- ou sous-représentation de chaque acide aminé en chaque position. Cette PSSM est ensuite utilisée pour une nouvelle itération dans le but de rechercher de nouvelles séquences similaires. Cette stratégie permet d'identifier des séquences homologues

ayant des relations de parenté plus lointaines. Un avantage de PSI-BLAST est qu'il couple chaque séquence proposée à un score de similarité par rapport à la séquence cible et à une mesure statistique (la *e-value*). Cette *e-value* mesure le nombre de séquences qu'il est possible de trouver par hasard dans la banque et ayant un score de similarité supérieur ou égal au score obtenu. Dans notre étude, quatre itérations de recherche sont réalisées et les séquences sélectionnées ont une *e-value* meilleures que 10^{-4}. La PSSM finale est conservée. Le logiciel blastpgp v2.2.13 a été utilisé (ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.13).

De façon similaire à l'étude précédente, chaque fragment de séquence à prédire fait 21 résidus de long (fragment d'intérêt de 11 résidus + extension de 5 résidus de part et d'autre pour tenir compte de l'environnement). Ainsi, pour chaque protéine cible, la PSSM obtenue est découpée en matrices chevauchantes de dimensions 20x21. Chaque fenêtre de séquence de 21 résidus est donc finalement représentée par une matrice de dimensions 20x21 caractérisant les spécificités de séquences observées parmi les séquences homologues (voir Figure 46). Finalement, les valeurs au sein des PSSMs sont normalisées pour être comprises dans l'intervalle [-1 ; +1] comme conseillé par (Chang and Lin 2001).

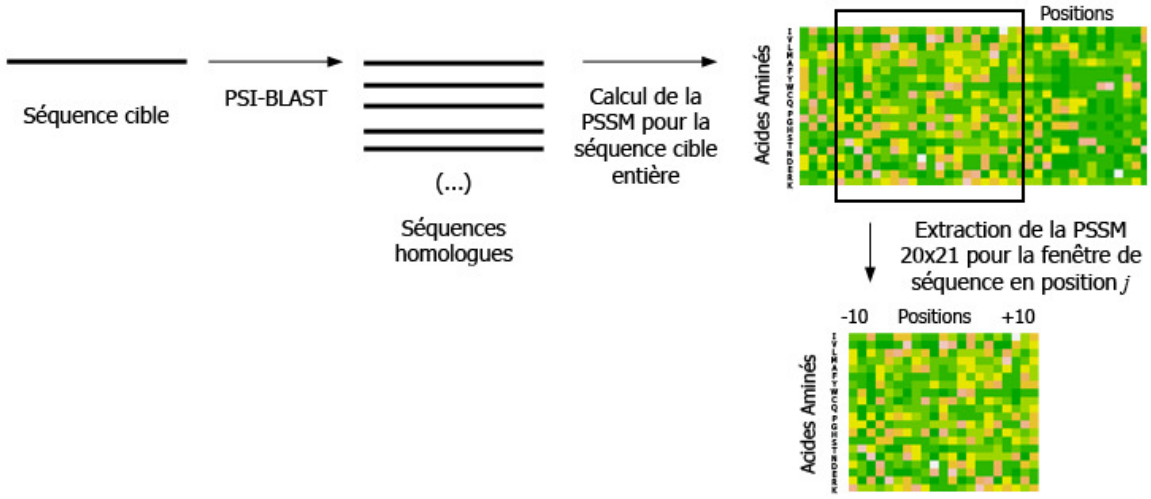

Figure 46. Enrichissement des fenêtres de séquence à prédire par des données évolutionnaires.

Des protéines homologues à la séquence cible sont tout d'abord recherchées grâce au logiciel PSI-BLAST. Ce dernier permet également le calcul d'une matrice de score position-spécifiques ou PSSM caractérisant les spécificités de séquence observées parmi les homologues. Pour chaque fenêtre de séquence à prédire (en position *j* et de longueur 21), une *sous*-PSSM de dimension 20x21 est extraite. Cette *sous*-PSSM décrira la fenêtre de séquence *j* utilisée pour l'apprentissage des experts et la prédiction.

4.1.2.3.2 Définition des experts par Machines à Vecteurs Supports (SVMs)

La seconde stratégie utilisée pour tenter d'améliorer la prédiction des structures locales, est la définition des experts par SVMs. Les SVMs correspondent à une généralisation des classifieurs linéaires (Hastie et al. 2001). Le principe d'apprentissage peut-être décomposé en deux étapes :

- Le jeu de données est tout d'abord projeté dans un espace de plus grande dimension en utilisant une fonction noyau. Cette fonction définit la similarité entre paires d'exemples au sein de cet espace (Lewis et al. 2006). Pour cette étude, nous avons choisi un noyau radial (*Radial Basis Function* en anglais ou RBF). Ce type de noyau a récemment été utilisé avec succès pour des prédictions de structures protéiques (Sander et al. 2006; Song et al. 2006). Il est défini comme suit :

  $K(x,x') = \exp(-\gamma\|x-x'\|^2)$ pour $\gamma > 0$, où x et x' sont deux exemples du jeu de données.
  Il implique le calibrage d'un paramètre γ.
- Par ailleurs, l'apprentissage des SVMs consiste à définir un hyperplan optimal situé le plus loin possible de tous les exemples d'une part et minimisant les erreurs d'apprentissage d'autre part. Cette procédure dépend d'un paramètre supplémentaire, *i.e.,* le paramètre C permettant de régler l'équilibre entre la minimisation des erreurs d'apprentissage et la maximisation des marges entre l'hyperplan et les exemples. De plus, un paramètre additionnel peut être optimisé : λ définit le poids des erreurs d'apprentissage sur les exemples positifs par rapport au poids des erreurs sur les exemples négatifs. Il correspond à une définition asymétrique du paramètre C.

Nous avons choisi d'utiliser le logiciel SVM^{light} adapté aux jeux de données de taille importante et possédant un algorithme rapide d'optimisation (Joachims 1999). Un expert SVM a été entrainé pour chaque classe structurale *s* en utilisant le sous-échantillon associé à *s* et tiré de l'échantillon 2 (paragraphes 3.3.4.1 et 4.1.2.2). Ainsi, des valeurs optimales pour γ, C et λ ont été calculées par grilles de validation croisée pour chaque classe *s* en fonction du pouvoir séparateur des experts (voir Figure 47). Pour conserver des temps de calculs raisonnables, ces paramètres ont été optimisés deux par deux, *i.e.* γ vs. C et γ vs. λ. Nous avons testé et adapté à notre procédure les intervalles de variation des paramètres conseillés par Hsu et collaborateurs (Hsu et al. 2003). Pour chaque classe *s* et pour chaque couple de paramètre testé (γ,λ) ou (γ,C), une procédure de validation croisée a été réalisée : l'échantillon d'apprentissage *s* est divisé aléatoirement en deux sous-échantillons contenant chacun le

même nombre d'exemples positifs et négatifs. Le premier sous-échantillon comprend 2/3 des données et est utilisé pour l'apprentissage du SVM avec le couple de paramètre testé. Le deuxième sous-échantillon est utilisé pour la validation. Pour chaque couple (γ,λ) ou (γ,C), cette procédure est répétée 5 fois et le taux moyen de bonne classification est calculé. Une fois le couple optimal sélectionné, l'apprentissage définitif du SVM est réalisé sur l'échantillon total dédié à la classe *s*.

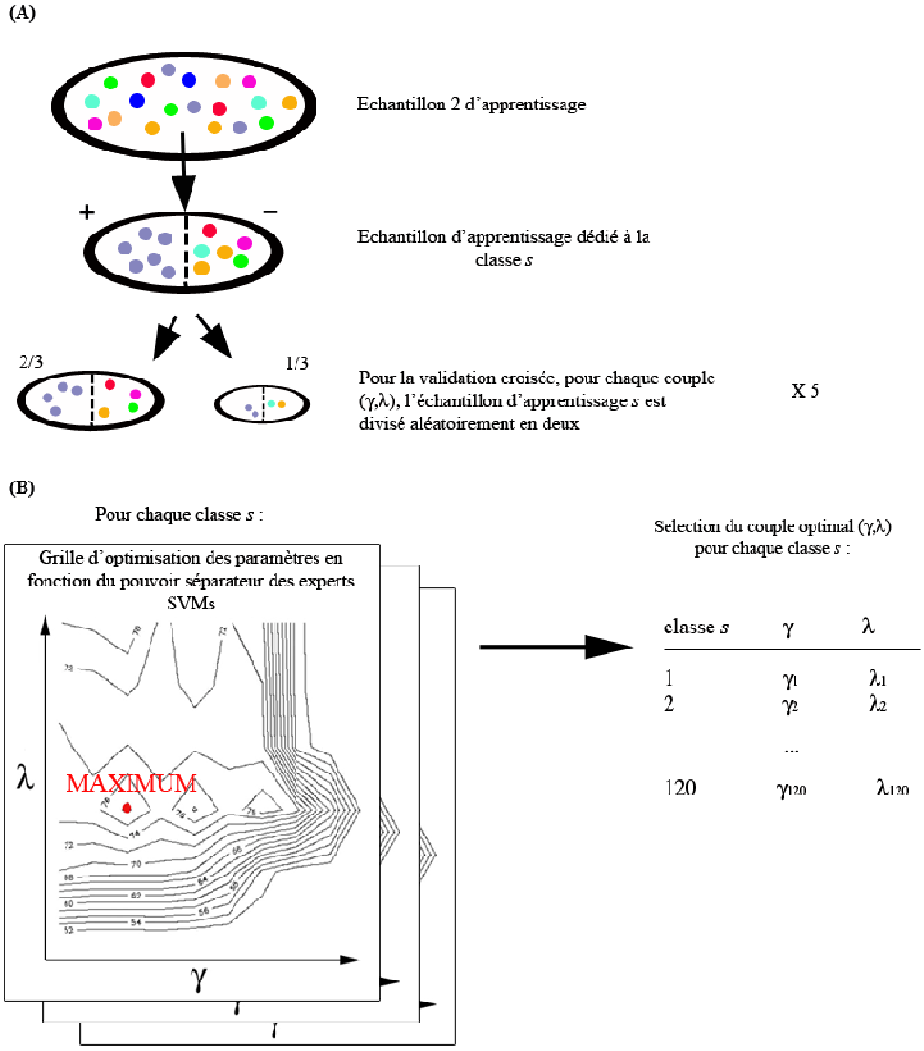

Figure 47. Calibrage des SVMs pour chaque classe structurale.

(A) Un échantillon d'apprentissage dédié à la classe *s* est extrait de l'échantillon 2. Il contient des fragments de séquence appartenant à sa classe (exemples positifs, bleus) et le même nombre fragments n'appartenant pas à sa classe (exemples négatifs). Dans le cadre de l'optimisation des paramètres du SVM de la classe *s*, une grille permettant de tester les performances associées à chaque couple de paramètre est réalisée. Ainsi, pour chaque couple, une validation croisée est réalisée. Dans ce but, l'échantillon d'apprentissage *s* est divisé en deux sous-échantillons : 2/3 des protéines sont réservés à l'apprentissage, 1/3 restent dédiés au calcul d'un taux de classification correcte. Cette procédure est réalisée 5 fois. Un taux de classification correcte moyen est alors calculé. (B) Un exemple de grille est présenté pour la classe *s*. Les taux moyens de classification correcte obtenus pour chaque couple de paramètre sont représentés par des courbes iso-contours. Pour chaque classe *s*, le couple de paramètres permettant d'obtenir le meilleur taux moyen est sélectionné.

Pour la prédiction d'une fenêtre de séquence *W*, l'expert SVM entraîné pour la classe *s* va calculer une valeur de décision v_s proportionnelle à la distance entre *W* et l'hyperplan optimisé pour *s*. v_s est positive si le SVM reconnaît la séquence *W* comme faisant partie des séquences associées à sa classe, négative sinon. Ainsi, la compatibilité d'un fragment de séquence *W* avec une classe structurale *s* est mesurée par la valeur de décision v_s du SVM ou *score*.

4.1.2.3.3 Sélection des candidats structuraux

Pour une prédiction donnée, les 120 scores obtenus (avec les experts définis par régression logistique ou SVM) sont analysés et les 5 PSLs ayant obtenus les meilleurs scores sont proposés en tant que candidats structuraux (cf. Figure 48).

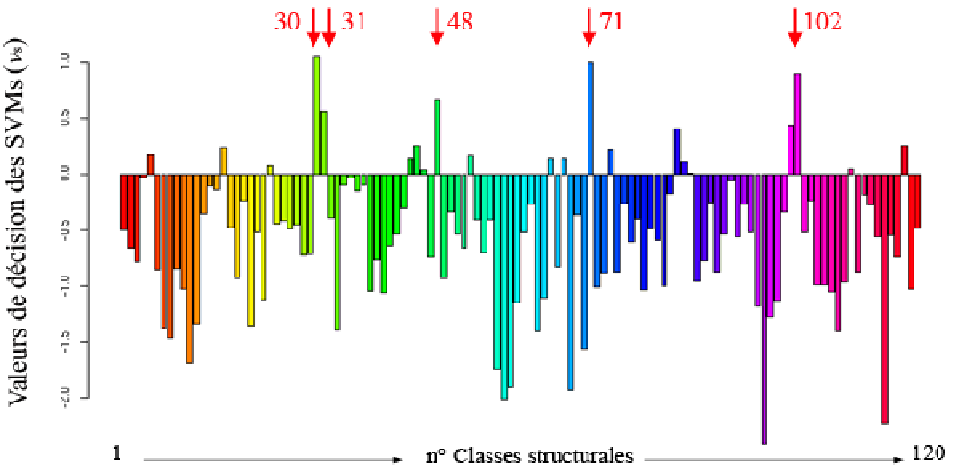

Figure 48. 120 valeurs de décision calculées par les experts SVMs lors de la prédiction d'une fenêtre de séquence.

Cette figure présente un exemple de prédiction obtenu avec les experts définis par SVM. Pour une fenêtre de séquence cible, les 120 experts SVMs ont calculé une valeur de décision v_s ou *score de compatibilité*. Les 5 PSLs prédits correspondent alors aux classes structurales ayant obtenus les meilleurs scores : dans l'ordre, les PSLs n° 30, 71, 102, 48 et 31.

4.1.2.4 Comparaison avec d'autres stratégies de prédictions des structures locales

Le paragraphe suivant (4.1.3) décrira les résultats obtenus en couplant les SVMs (paramétrés avec le couple (γ,λ)) aux données évolutionnaires. Cette stratégie, nommée *SVM_PSSM*, a en effet permis d'obtenir les meilleures performances de prédiction. Nous comparerons tout d'abord l'efficacité de cette prédiction avec :

i) Une prédiction aléatoire : cinq candidats sont choisis aléatoirement pour chaque fragment de séquence de l'échantillon 3 de validation. Ce type de prédiction servait déjà de référence dans l'étude précédente (Benros et al. 2006).

ii) Une prédiction dite *naïve* : pour chaque séquence de 21 résidus de long de l'échantillon 3, les 5 cinq séquences les plus similaires de l'échantillon 1 sont sélectionnées. Les scores de similarité sont calculés en utilisant la matrice BLOSUM 62. Les fragments de structure correspondants sont alors superposés à celui de la séquence cible, en considérant uniquement les 11 résidus centraux (positions -5 à +5 dans la fenêtre de séquence cible). Cette stratégie est similaire aux premières étapes de prédiction de modèles protéiques par assemblage de fragments comme dans ROSETTA (Rohl et al. 2004).

Des comparaisons avec les autres stratégies décrites dans le Tableau 8 ainsi qu'avec d'autres méthodes de pointe de prédictions des structures locales seront présentée en discussion.

4.1.3 Résultats

Ce paragraphe présente les résultats obtenus avec la stratégie la plus performante : *SVM_PSSM*.

4.1.3.1 Evaluation de la stratégie de prédiction

4.1.3.1.1 Evaluation globale des listes de structures locales candidates prédites

Le taux de prédictions correctes obtenu en considérant uniquement la classe structurale assignée (ou Q_{120}) est égal à 38,8 % (Tableau 9). Ce résultat correspond à un gain significatif de 34,6 % par rapport à une prédiction aléatoire et de 18,3 % par rapport à une prédiction *naïve* (paragraphe 4.1.2.4).

Selon l'évaluation basée sur un critère géométrique de 2,5 Å, le taux de prédiction obtenu est de 63,1 %. De même, les gains par rapport à des prédictions aléatoire et *naïve* sont importants : 38,0 % et 15,2 % respectivement. Le Tableau 10 (3ème colonne, première ligne) montre que pour les prédictions considérées comme correctes (*i.e.,* 63,1 % des fragments de l'échantillon de validation), au moins un des PSLs candidats propose une approximation à seulement 1,45 Å de la vraie structure locale en moyenne. De même, pour ces fragments bien prédits, l'approximation moyenne fournie par les 5 candidats est de 2,54 Å. Enfin, si tous les

fragments de l'échantillon de validation sont considérés, une approximation minimale satisfaisante de 2,09 Å est toujours disponible en moyenne parmi les 5 candidats. L'approximation moyenne donnée par les 5 candidats est alors de 3,03 Å.

Tableau 9. Résultats de prédictions des structures locales.

Analysis of the structural prediction results						
Experts definitions and target sequence window representation		*LR_seq*	*SVM _seq*	*SVM _PSSM*	*SVM _PSSM gains over random*	*SVM_PSSM gains over Similar Sequences Search*
Proportion of true positives (%)		31.43	30.61	38.75	34.56	18.34
Prediction rate (%) (approximation < 2.5Å)		**55.48**	**55.54**	**63.13**	**37.95**	**15.18**
Results per secondary structures categories (%)						
H	*Proportion of true positives*	40.06	39.10	50.68	46.54	23.47
	Prediction rate (approximation < 2.5Å)	**76.80**	**75.45**	**84.60**	**43.01**	**3.46**
E	*Proportion of true positives*	20.95	25.45	34.08	30.19	18.82
	Prediction rate (approximation < 2.5Å)	**57.43**	**63.18**	**73.03**	**39.26**	**23.94**
C	*Proportion of true positives*	33.49	31.46	38.03	33.72	17.31
	Prediction rate (approximation < 2.5Å)	**45.06**	**43.85**	**49.47**	**36.52**	**16.17**
Ext	*Proportion of true positives*	22.89	21.80	28.35	24.16	13.97
	Prediction rate (approximation < 2.5Å)	**47.10**	**48.31**	**56.30**	**33.88**	**23.22**

Les résultats de 3 schémas de prédiction sont présentés : *LR_seq* (experts définis par régression logistique, séquence seule), *SVM_seq* (experts définis par SVM, séquence seule) et *SVM_PSSM*. Les deux dernières colonnes quantifient les gains de la stratégie SVM_PSSM par rapport à des prédictions aléatoire et *naïve*. Le nombre de candidats par fenêtre de séquence est fixé à 5. Tableau extrait de (Bornot et al. 2009).

Tableau 10. Approximation structurale fournie par la prédiction des structures locales.

Average geometrical approximation of the local structure prediction (A)

		All predicted fragments		Fragments correctly predicted according to the geometrical criteria (<2.5 Å)	
		Minimal RMSD[a]	Mean RMSD[a]	Minimal RMSD[a]	Mean RMSD[a]
Set 3		2.09	3.03	1.45	2.54
Secondary structures categories	H	1.21	2.28	0.83	1.95
	E	2.17	3.14	1.78	2.77
	C	2.48	3.41	1.75	2.93
	Ext	2.41	3.20	1.93	2.86

[a]Over the five candidates per fragment.

Les résultats de prédiction globaux sur l'échantillon 3 de validation, sont présentés en première ligne. Puis, les classes structurales sont présentées en fonction de leurs catégories proches des structures secondaires : H correspond aux structures hélicoïdales, E aux structures étendues, C aux structures de connexion et Ext aux extrémités de structures étendues. Tableau adapté de (Bornot et al. 2009).

Il est important de noter que la prédiction n'est pas biaisée vers les classes les plus fréquentes et les plus homogènes. La prédiction de chacune des 120 classes est largement meilleure que l'aléatoire. Ainsi, le taux de vrais positifs par classe atteint en moyenne 33,7 %, soit un gain de 30,4 % en moyenne par rapport à une prédiction aléatoire. Selon l'évaluation basée sur un critère géométrique de 2,5 Å, le taux de prédiction par classe atteint 58,1 % en moyenne. Ce résultat représente un gain moyen de 39,2 % par rapport à une prédiction aléatoire. Ces gains sont bien équilibrés et concernent toutes les classes structurales. Le gain le plus faible est observé pour la classe n° 113 mais est toujours supérieur de 13,3 points par rapport à une prédiction aléatoire. Cette classe est celle qui présente la plus large variabilité structurale avec un Cα RMSD *intra*-classe de 2,44 Å en moyenne (voir paragraphe 3.3.2).

4.1.3.1.2 Evaluation de la prédiction en catégories de PSLs proches des structures secondaires

Comme réalisé dans l'étude précédente menée par Benros et collaborateurs, la prédiction peut également être évaluée en fonction de catégories de PSLs proches des structures secondaires (Benros et al. 2006). Les taux de prédictions correctes obtenus en considérant uniquement la classe structurale assignée varient alors entre 28,4 % pour les PSLs de type étendu et 50,9 % pour les PSLs de type hélicoïdal. Ces résultats correspondent à des gains de 24,2 % et 46,5 % respectivement en comparaison d'une prédiction aléatoire et à des gains de 14,0 % et 23,5 % par rapport à une prédiction *naïve* (Tableau 9).

Les taux de prédiction reposant sur le critère géométrique de 2,5 Å, varient entre 49,5 % et 84,6 % (Tableau 9). Ainsi, par rapport à une prédiction aléatoire, des gains bien équilibrés varient entre 43,0 et 33,9 %. En comparaison d'une prédiction *naïve*, les gains restent élevés, *i.e.,* respectivement 23,9 %, 16,2 % et 23,2 % pour les classes structurales étendues, de connexion et d'extrémités de structures étendues. Une exception toutefois concerne les PSLs hélicoïdaux : un gain de seulement 3,5 % est observé par rapport à une prédiction *naïve*. Cette performance moins significative peut être attribuée à deux facteurs : *(i)* les fortes spécificités de séquence associées aux hélices rendent probablement plus aisée une prédiction *naïve* basée uniquement sur les similarités de séquences, *(ii)* d'autre part, le critère géométrique de 2,5 Å n'est généralement pas suffisamment strict pour évaluer précisément la prédiction des structures locales hélicoïdales. Selon le critère géométrique à 1,5 Å (voir Annexe 1), notre stratégie de prédiction *SVM_PSSM* permet d'obtenir un taux de prédictions correctes de 67,8 % pour les structures hélicoïdales. Dans ce cadre, un gain de 6,5 points est observé par rapport à une prédiction *naïve*.

Par ailleurs, pour les 63,1 % de fenêtres de séquence considérées comme bien prédites (critère géométrique de 2,5 Å), la meilleure approximation disponible parmi les 5 PSLs candidats atteint en moyenne 0,83, 1,78, 1,75 et 1,93 Å pour les structures hélicoïdales, étendues, de connexion et d'extrémités de structures étendues respectivement (Tableau 10). De même, pour l'ensemble des fragments (bien prédits ou non), des approximations de 1,21, 2,17, 2,48 et 2,41 Å respectivement sont toujours disponibles en moyenne parmi les 5 candidats. Ces résultats sont particulièrement significatifs. En effet, les variabilités structurales *intra*-classes (dans les classes assignées) sont en moyenne de 1,29 Å (σ = 0,98), 2,14 Å (σ = 0,48), 3,34 Å (σ = 0,61) pour les structures hélicoïdales, étendues, de connexion respectivement. La variabilité des extrémités de structures étendues peut être décomposée en deux : 2,87 Å (σ = 0,68) and 2,58 Å (σ= 0,53) pour les entrées et les sorties de brin respectivement (Benros 2005). Nos prédictions sont donc en moyenne en deçà des variabilités *intra*-classes.

4.1.3.2 Exemples de prédictions

4.1.3.2.1 Exemples de prédiction sur la protéine ARL3-GDP de la classe SCOP α/β

La Figure 49 présente cinq exemples de prédictions pour la protéine de liaison au calcium de code PDB 1FZQ (Hillig et al. 2000). Cette protéine de la classe SCOP des protéines *α/β* (Murzin et al. 1995), fait 175 résidus de long. Le taux de prédictions correctes obtenu pour cette protéine est de 72,4 % avec la stratégie *SVM_PSSM* (au lieu de 55,6 % avec la stratégie initiale).

Les deux premiers exemples (a) et (b) correspondent respectivement à un cœur et une extrémité C-terminale d'hélice. Dans les deux cas, le candidat du premier rang est également le PSL qui avait été assigné à partir de la structure cristallographique. La meilleure approximation possible de la vraie structure locale a donc été obtenue.

Le troisième exemple (c) concerne une structure de connexion. Le candidat du premier rang est à 3,9 Å de la structure locale réelle (Cα RMSD). Le PSL assigné, n° 14, à 2,2 Å de la structure cristallographique, n'est cette fois pas retrouvé parmi les 5 candidats. Toutefois, une approximation de 2,7 Å est obtenue grâce au candidat du second rang, le PSL 88. Ce dernier adopte une forme globale et une orientation C-terminale très similaire au fragment réel. Il est important de souligner que cette prédiction, bien que tout à fait satisfaisante qualitativement, ne fournit pas d'approximation meilleure que notre critère géométrique de 2,5 Å utilisé pour l'évaluation de la prédiction. Ainsi, cette prédiction fait partie des prédictions considérées

comme incorrectes. Cet exemple illustre le caractère strict du critère géométrique de 2,5 Å pour les structures de connexion notamment.

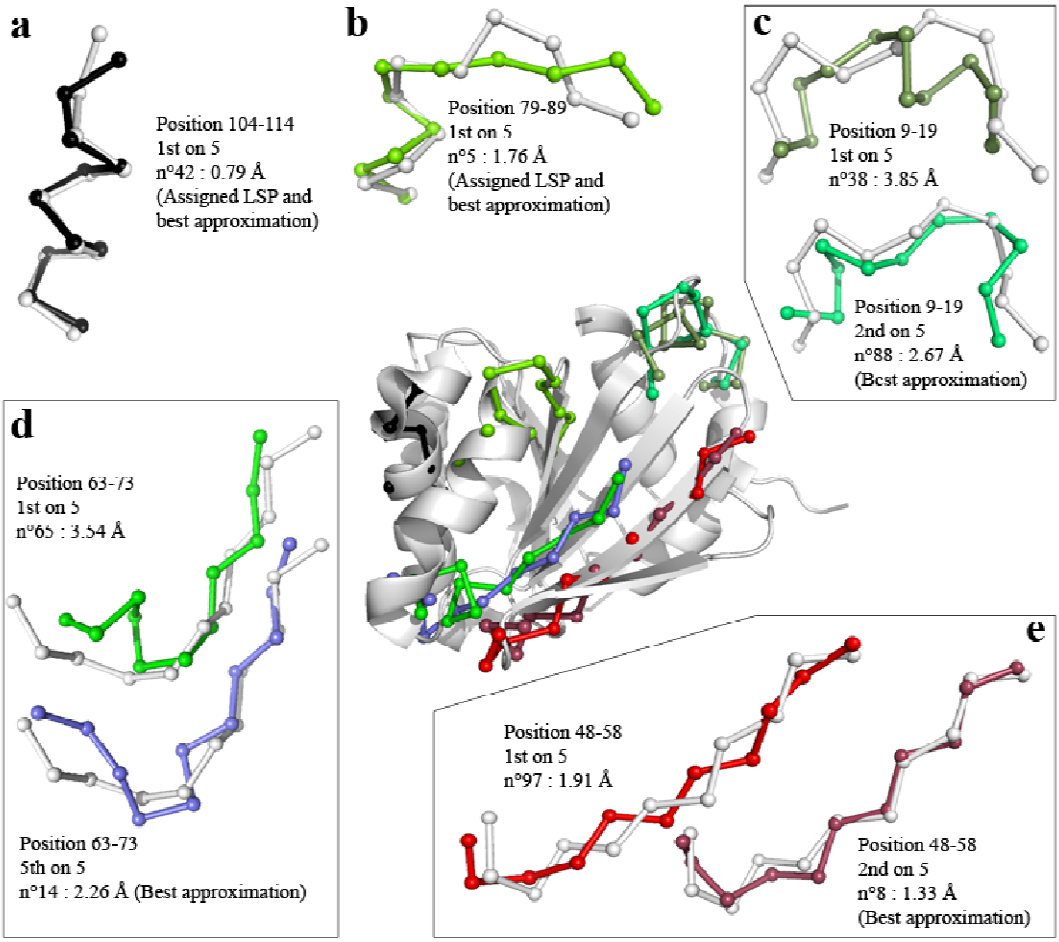

Figure 49. Cinq exemples de prédiction sur une protéine de classe SCOP alpha/béta.

La structure réelle est en blanc. Les PSLs prédits sont en couleur : les structures hélicoïdales, étendues, de connexion et d'extrémités de structure étendues sont en noir, rouge, vert et bleu respectivement. Pour chaque prédiction sont indiqués : la position de la fenêtre de séquence dans la protéine, le rang du PSL prédit, le numéro de sa classe structurale et l'approximation qu'il fournit par rapport à la vraie structure locale (Cα RMSD). Figure extraite de l'annexe 7 de (Bornot et al. 2009).

Les deux derniers exemples concernent la prédiction d'une extrémité de structure étendue et d'une structure étendue (exemples (d) et (e)).

Pour l'extrémité de structure étendue (d), le candidat ayant obtenu le meilleur score est le PSL 64. Ce dernier est assez éloigné de la vraie structure locale (Cα RMSD = 3,5 Å). Toutefois, le dernier des 5 candidats, le PSL 14, permet une approximation à 2,3 Å du repliement réel. Cette approximation est très proche de la meilleure qu'il était possible d'obtenir : le PSL assigné 110 est à 2,2 Å du vrai fragment.

Dans le cas du cœur de structure étendue, le candidat du premier rang, le PSL 97, fournit une approximation à 1,9 Å et une bonne orientation de l'extrémité N-terminale du fragment. Le candidat du second rang fournit la meilleure approximation (Cα RMSD = 1,33 Å). Parmi les 5 candidats proposés, 4 permettent une approximation correcte du fragment, bien que le PSL assigné 96 (avec un Cα RMSD = 1,05 Å) ne soit pas proposé.

4.1.3.2.2 Exemples de prédiction sur une protéine de liaison aux odeurs de classe SCOP tout-β

La Figure 50 présente cinq autres exemples de prédictions pour la protéine de liaison aux odeurs (code PDB 1OBP) (Bianchet et al. 1996) de la classe SCOP *tout-β*. Le taux de prédictions correctes obtenu pour cette protéine est de 74,5 % avec la stratégie *SVM_PSSM* (au lieu de 65,7 % avec la stratégie initiale).

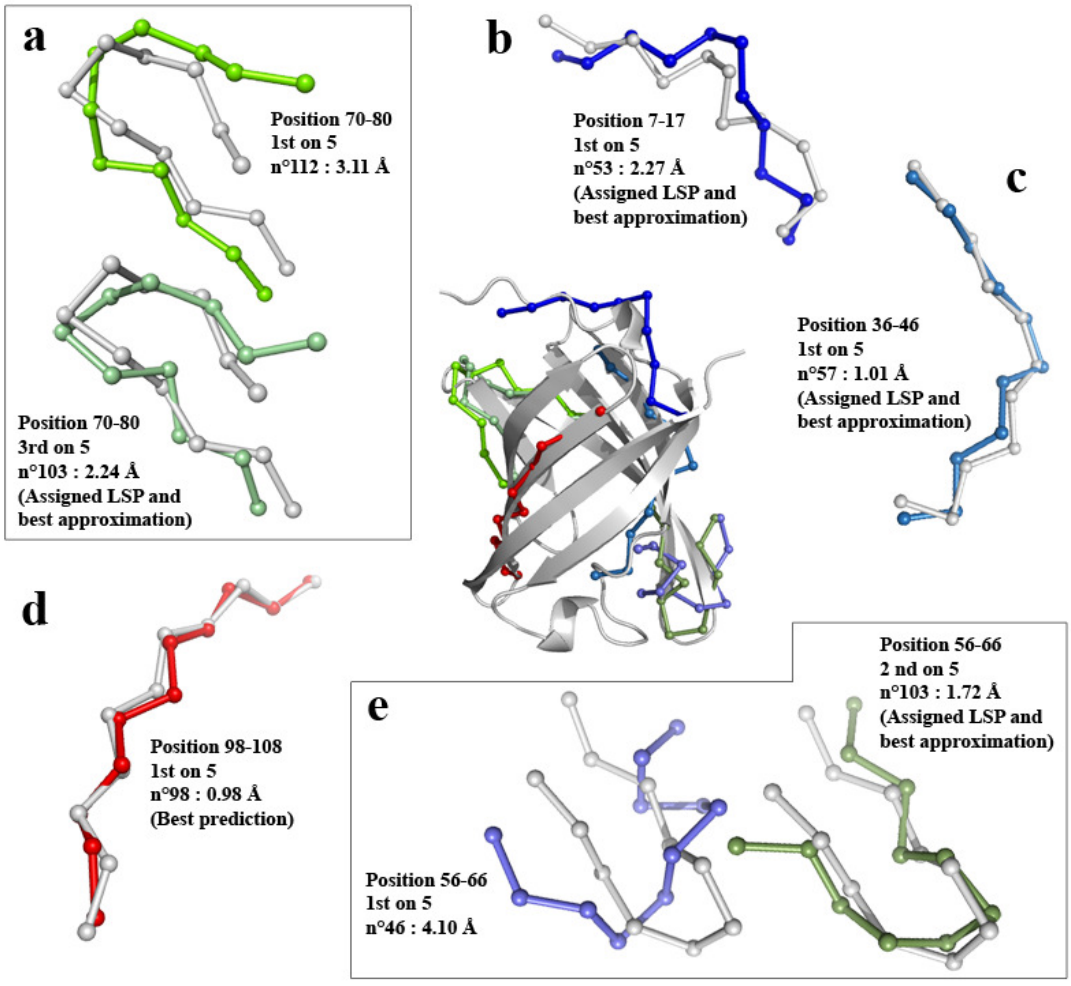

Figure 50. Cinq exemples de prédiction sur une protéine de classe SCOP Tout béta.

Voir la légende de la Figure 49. Figure extraite de l'annexe 5 de (Bornot et al. 2009).

Les exemples (a) et (e) sont très similaires. Les fragments à prédire font partie d'une structure super-secondaire en β-hairpin incluant un β-turn de type IV au niveau de la boucle (selon Promotif (Hutchinson and Thornton 1996) dans la PDBsum (Laskowski et al. 2005a)). Dans les deux cas, le PSL assigné n° 103 est retrouvé parmi la liste des candidats prédits et permet une très bonne approximation à 2,2 Å et à 1,7 Å pour les exemples (a) et (e) respectivement.
Les exemples (b) et (c) correspondent à l'extrémité N-terminale de la protéine et à une extrémité N-terminale de brin β. Pour ces exemples, le candidat ayant obtenu le meilleur score de compatibilité est aussi le PSL qui avait été assigné à partir de la structure cristallographique.
Dans l'exemple (d), le PSL 10 assigné n'est pas retrouvé parmi les candidats prédits. Toutefois, le candidat du premier rang, le PSL 98, donne une très bonne approximation de la structure réelle (Cα RMSD = 1 Å). La courbure globale du brin est notamment bien prédite.

4.1.3.2.3 Exemples de prédiction sur une protéine de liaison au calcium de la classe SCOP tout-α

Enfin, la Figure 51 présente cinq exemples de prédictions sur une protéine *tout-α* : la protéine de liaison au calcium de code PDB 1K94 (Jia et al. 2001). Le taux de prédictions correctes obtenu pour cette protéine est de 71,5 % avec la stratégie *SVM_PSSM* (au lieu de 53,6 % avec la stratégie initiale).
Les exemples proposés ici concernent majoritairement des structures de connexion, les plus difficiles à prédire.
Pour les exemples (a), (b) et (e), le PSL assigné à partir de la vraie structure est retrouvé dans la liste de candidat. Ainsi, dans l'exemple (a) correspondant aux 9 derniers résidus d'une boucle longue suivie de l'extrémité N-terminale d'une hélice, une approximation très satisfaisante de 2,2 Å est fournie par le candidat du premier rang. Dans l'exemple (b) incluant une boucle de 6 résidus de long entre 2 hélices, le candidat du second rang fournit une approximation remarquable de 1,6 Å. Enfin, dans l'exemple (e), un fragment comprenant une boucle de 9 résidus incluant un γ-turn puis un β-turn de type I, le 4ème candidat fournit une approximation à 2,4 Å.
L'exemple (c) comprend une boucle de 10 résidus de long incluant un β-turn de type II'. Le prototype assigné n'est pas retrouvé parmi les candidats et la meilleure approximation obtenue est à 2,8 Å de la structure locale réelle. Cette prédiction est donc jugée incorrecte selon le critère géométrique à 2,5 Å. Pourtant, comme dans l'exemple (c) de la Figure 49, la forme générale du meilleur fragment prédit n'est pas si différente du vrai repliement. Les

orientations des extrémités de la structure locale sont de plus assez satisfaisantes. Ainsi, ces prédictions jugées incorrectes mais fournissant des approximations qualitativement satisfaisantes pourraient bien être *suffisamment* précises dans le cadre de la construction de modèles protéiques par assemblage de fragments. Nous verrons dans la section 6 que la prise en compte de la flexibilité des structures soutient cette hypothèse.

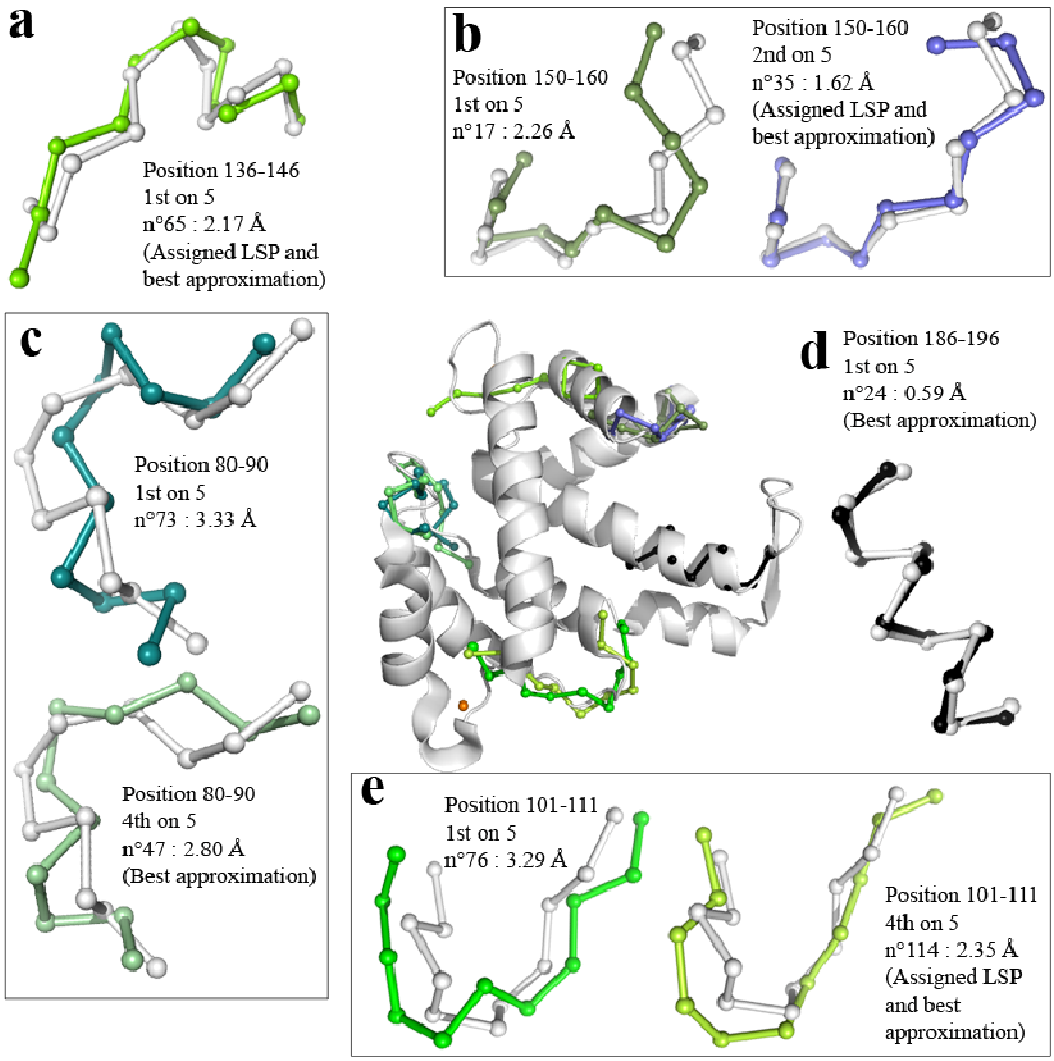

Figure 51. Cinq exemples de prédiction sur une protéine de classe SCOP Tout alpha.
Voir la légende de la Figure 49. Figure extraite de l'annexe 6 de (Bornot et al. 2009).

4.1.4 Discussion – Conclusion

Dans ce paragraphe, je comparerai la pertinence et les résultats de la stratégie *SVM_PSSM* à d'autres schémas de prédiction plus sophistiqués mais aussi à des méthodes de pointe publiées récemment par d'autres laboratoires.

4.1.4.1 Comparaison avec d'autres stratégies de prédiction

Comme présenté dans le Tableau 8, outre la stratégie *SVM_PSSM*, trois autres méthodes de prédiction ont été testées :

- *SVM_seq*: les SVMs sont entrainés à partir de la séquence seule décrite comme dans (Benros et al. 2006).
- *LR_seq*: les experts sont définis par régression logistique à partir de la séquence seule. *LR_seq* est une réévaluation de la méthode développée par Benros et collaborateurs lorsque 5 candidats sont prédits quelque soit leur position dans la séquence. En effet, la méthode initiale prévoyait un nombre variable de candidats pour chaque position, *i.e.* 5 candidats maximum pouvaient être prédits (voir paragraphe 3.3.4.1).
- *LR_PSSM*: définition des experts par régression logistique et utilisation des PSSMs pour décrire les fragments de séquence.

L'annexe 1 présente de façon détaillée les résultats obtenus avec ces différentes méthodes.

4.1.4.1.1 Performances similaires des experts définis par SVM ou régression logistique sur la séquence seule

Le couplage des experts définis par SVM avec l'utilisation de la séquence seule (*SVM_seq*) mène à un Q_{120} de 30,6 % (voir Tableau 9). Ce résultat est très similaire à celui obtenu avec la stratégie initiale *LR_seq* reposant sur la régression logistique. De même, selon le critère géométrique à 2,5 Å, les taux de prédiction obtenus avec *SVM_seq* et *LR_seq* sont identiques (55,5 %). Les taux par catégories de PSLs sont aussi identiques (voir Tableau 9 et Annexe 1). Ainsi, les experts définis par SVM ou régression logistique montrent des performances équivalentes pour tirer partie de l'information codée dans la séquence seule.

4.1.4.1.2 Amélioration de la prédiction obtenue en couplant les SVMs aux données évolutionnaires

En revanche, l'utilisation d'informations évolutionnaires avec les SVMs (*SVM_PSSM*) mène à une amélioration significative des performances de prédiction par rapport à *LR_seq* et *SVM_seq*. En effet, le Q_{120} et le taux de prédiction basé sur le critère géométrique gagnent tous deux plus de 7 points. Des analyses plus détaillées montrent également que la prédiction est améliorée pour les 4 catégories de PSLs basées sur les structures secondaires. Les gains

vont de 4,4 % pour les structures de connexion à 15,6 % pour les structures étendues (cf. Annexe 1).

De plus, la majorité des 120 classes structurales bénéficient de cette amélioration significative du taux de prédiction par rapport à la méthode originale. En effet, le Q_{120} de 74,2 % des classes est augmenté. Par ailleurs, la corrélation entre l'augmentation du taux de prédiction et la fréquence d'apparition des classes est très faible ($r_{Pearson}$ = 0,21) : cette amélioration n'est donc pas biaisée vers les classes fréquentes. Ainsi, un gain est également observé pour la prédiction des 2/3 des 35 classes les moins peuplées (rassemblant moins de 5 % des structures locales). La moitié d'entre elles gagnent plus de 10 points. De même, selon le critère géométrique à 2,5 Å, le taux de prédiction de 88,3 % des classes est meilleur que celui obtenu avec la méthode initiale *LR_Seq* (cf. Figure 52). Seules 12 classes de structures de connexion et d'extrémités de structures étendues sont moins bien prédites. Elles représentent uniquement 7,4 % des fragments.

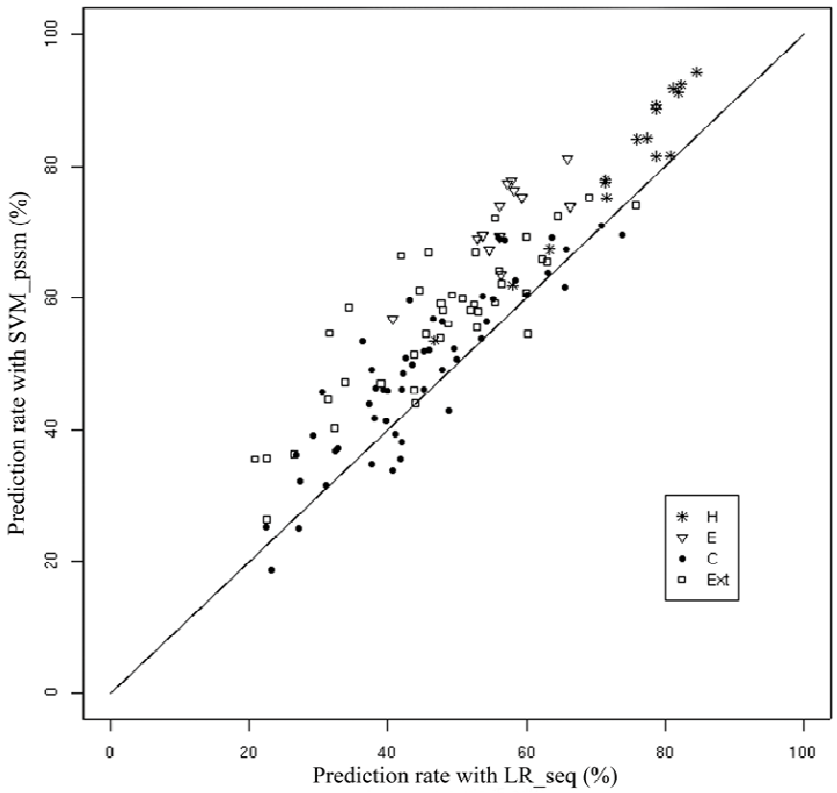

Figure 52. Amélioration globale de la prédiction des structures locales obtenue en couplant les SVMs à des informations évolutionnaires.

Les taux de prédiction obtenus avec la stratégie *SVM_PSSM* pour les 120 classes sont comparés avec les taux obtenus avec *LR_seq*. Les taux sont évalués en fonction du critère géométrique à 2,5 Å. Les classes structurales sont représentées en fonction de leurs catégories proches des structures secondaires : H correspond aux structures hélicoïdales, E aux structures étendues, C aux structures de connexion et Ext aux extrémités de structures étendues. Figure extraite de (Bornot et al. 2009).

Finalement, 85,3 % des protéines bénéficient d'une amélioration de leur prédiction.

4.1.4.1.3 Couplage des informations évolutionnaires avec la régression logistique

Dans un but de comparaison, nous avons également étudié les performances du couplage information évolutionnaires (PSSM) / régression logistique (LR). De manière surprenante, l'utilisation de cette stratégie (*LR_PSSM*) est associée à une chute spectaculaire des résultats de prédiction. Différentes explications peuvent être avancées pour expliquer ce phénomène. La première est la taille de la banque de données. Pour certaines classes, le nombre d'exemples d'apprentissage est trop faible par rapport au nombre important de variables explicatives, pour permettre un calcul correct des coefficients des fonctions logistiques, *i.e.,* pour un fragment, un PSSM est de dimension 20x21 soit 420 variables explicatives. Toutefois, l'explication principale de cette chute de performance réside dans le calcul des scores de compatibilité *séquence cible / classe structurale* réalisé par les experts. Lorsque ces derniers sont définis par régression logistique ces scores sont des probabilités (voir paragraphe 3.3.4.1). Or, une analyse détaillée montre que la distribution des probabilités n'est par uniformément répartie entre 0 et 1, mais confinée au niveau des valeurs extrêmes. Ainsi, face à cette réponse presque binaire des experts, le choix, par le jury, des PSLs candidats selon leur score de compatibilité avec la séquence n'est plus approprié et mène à des prédictions inadaptées (cf. Annexe 1).
Ces résultats renforcent donc la pertinence du couplage des SVMs avec l'utilisation des PSSMs pour la description des séquences cibles.

4.1.4.2 Comparaison avec des méthodes de prédiction de pointe

La diversité des méthodes de prédiction des structures locales existantes rendent leur comparaison directe particulièrement difficile. En effet, elles reposent le plus souvent sur différentes représentations des structures locales et différents protocoles de prédiction et d'évaluation. Nous avons donc adapté nos résultats pour permettre une comparaison avec trois types de prédictions des structures locales :
(i) En se basant sur l'étude réalisée précédemment, il est possible de comparer nos résultats à des méthodes de prédiction des angles de torsion du squelette polypeptidique (Benros et al. 2006). Comme nous avons vu dans le paragraphe 3.3.4.3, la méthode de prédiction initiale, avec 4,2 candidats en moyenne par séquence cible, donnait déjà des résultats tout à fait comparable avec les méthodes développées par Bystroff (Bystroff et al.

2000), Kuang (Kuang et al. 2004a) ou Yang et Wang (Yang and Wang 2003). Or, notre nouvelle stratégie *SVM_PSSM* est significativement plus performante que la méthode originale (+ 12% environ) (tableau de l'Annexe 1, 9^{e} colonne, 2^{e} ligne). Par conséquent, il est possible de considérer que notre approche est plus que compétitive avec les méthodes déjà existantes spécialisées dans la prédiction des angles de torsion.

Par ailleurs,

(ii) Une comparaison a également été effectuée avec une méthode associée à un autre alphabet structural.

(iii) Enfin, nous avons évalué les performances de notre approche par rapport à des méthodes actuelles de modélisation 3D de boucles longues.

Je développerai dans les deux paragraphes qui suivent la comparaison entre notre stratégie *SVM_PSSM* et ces deux derniers types de prédiction.

4.1.4.2.1 Prédictions associées à des alphabets structuraux

Sander et collaborateurs ont défini un nouvel alphabet structural associé à une méthode de prédiction des structures locales (Sander et al. 2006) (paragraphe 3.1.2.2). 27 structures locales de 7 résidus de long ont été obtenues par discrétisation de l'espace des séquences et des structures. Plusieurs stratégies de prédiction ont été testées. La méthode sélectionnée par les auteurs repose sur la combinaison de Machines à Vecteurs Supports et d'une représentation des séquences cibles sous forme de profils basés sur les propriétés physico-chimiques des résidus.

Une comparaison avec la stratégie *SVM_PSSM* est ardue car le nombre de classes structurales et la longueur des fragments prédits sont très différents. Toutefois, dans le but de nous rapprocher des conditions de prédiction de Sander et associés, nous avons regroupé nos 120 PSLs en 27 classes en fonction de leur similarité structurale. Le Tableau 11 présente les taux de prédiction, Q_{27}, obtenus par Sander et collaborateurs pour des structures locales de 7 résidus de long (Sander et al. 2006). Ce taux varie entre 34 et 64 % pour 1 à 5 candidats par fenêtre de séquences prédites. Ces résultats sont comparables à ceux obtenus avec notre stratégie *SVM_PSSM* : le Q_{27} varie également de 32 à 61 %. Ces taux sont d'autant plus satisfaisants que nous prédisons des structures de 11 résidus de long, *i.e.,* 4 résidus de plus que celles prédites par Sander.

Tableau 11. Comparaison avec la méthode de prédiction des structures locales de Sander et associés.

	True positive rate for the top-*k* ranked predicitions (%)		
	Property profile+RF	SVM_PSSM	
Rank *k*	7-residue local structures 27 classes	11-residue local structures 120 classes	11-residue local structures 27 classes
1	34	13.56	31.84
2	46	22.24	44.08
3	54	28.91	51.6
4	60	34.22	57.05
5	64	38.75	61.42

La 2ème colonne correspond à la prédiction de *Sander et al.*, la 3ème colonne correspond à notre prédiction et la 4ème à notre prédiction en 27 classes. Tableau extrait de (Bornot et al. 2009).

Les performances de notre stratégie peuvent également être évaluées par l'analyse de courbes ROC (*Receiver Operating Characteristic*) construites pour chaque classe structurale. Ces courbes représentent l'évolution du taux de vrai positif en fonction du taux de faux positif (Fawcett 2003). Elles sont indépendantes de la distribution des classes et peuvent être calculées pour des méthodes de prédiction fournissant un score continu. Ainsi, pour chaque fragment, la différence entre le score du candidat du premier rang et le score de chaque classe a été calculée. Pour une classe donnée, cette différence quantifie la performance de l'expert associé et la replace dans le cadre de la prédiction globale. Cette performance est ensuite visualisée grâce à l'aire sous la courbe ROC (*Area Under the ROC curve* en anglais ou AUC). L'AUC peut varier de 0,5 (pour une prédiction aléatoire) à 1 (pour une prédiction parfaite). Avec la méthode *SVM_PSSM*, les AUCs obtenues varient entre 0,71 et 0,92 avec une moyenne à 0,82 (cf. Figure 53). Les valeurs minimales et maximales correspondent respectivement aux classes 55 (structure de connexions) et 33 (extrémités de structures étendues). A titre de comparaison, l'approche de Sander et collaborateurs mène à des AUCs variant de 0,68 à 0,88. Finalement, cette évaluation de notre stratégie *via* des courbes ROC contribue à mettre en évidence les performances au moins équivalentes de *SVM_PSSM* par rapport à la méthode de Sander et collaborateurs, mais utilisant des fragments protéiques plus longs.

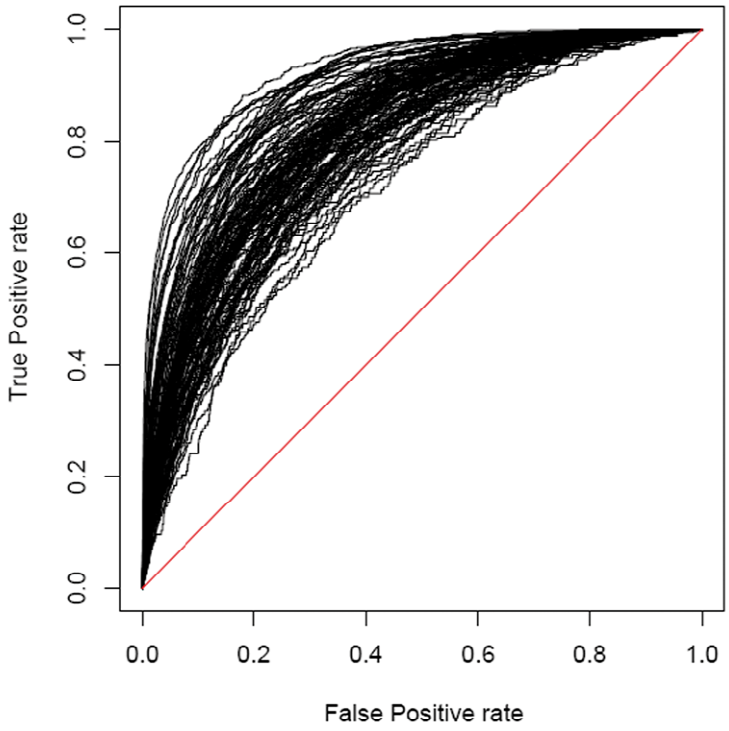

Figure 53. Courbes ROC pour la prédiction des 120 classes de structures locales.
Figure extraite de (Bornot et al. 2009).

4.1.4.2.2 Prédiction des boucles longues

La prédiction des boucles est une étape majeure et délicate de la modélisation par homologie de structures protéiques (voir paragraphe 2.3.8.1). Comme nous l'avons vu dans le paragraphe 3.2.5.1, cette prédiction est fréquemment réalisée *a posteriori* de la modélisation de l'arrangement des structures secondaires répétitives. Ainsi, les différentes approches de prédiction des boucles bénéficient de la connaissance partielle ou complète du reste de la structure.

Nous avons analysé l'efficacité de notre stratégie *SVM_PSSM* pour fournir une prédiction rapide et satisfaisante de la conformation d'une région de boucle. Dans ce but, nous avons comparé la précision obtenue par des méthodes de prédiction *ab initio* dédiées aux boucles à la précision obtenue avec *SVM_PSSM*. L'analyse a donc été effectuée uniquement sur les fragments assignés à des PSLs caractérisant des structures de connexion. Des exemples de prédiction de structure de connexion sont disponibles dans les Figure 49, Figure 50 et Figure 51.

Actuellement, les protocoles de modélisation des boucles reposent à la fois sur l'échantillonnage des conformations possibles et sur l'attribution d'un score à ces dernières (voir paragraphe 2.3.8.1). Ce score a pour but de permettre la sélection des conformations les

plus favorables énergétiquement. Ainsi, la prédiction des boucles longues de plus de 10 résidus reste encore un défi car l'échantillonnage requiert alors des capacités de calcul très importantes (Zhu et al. 2006).

Nous nous sommes concentrés sur cinq méthodes de prédiction des boucles récemment comparées les unes par rapport aux autres : *LoopBuilder* (Soto et al. 2008), la méthode de prédiction des boucles associée à *Modeller* (Fiser et al. 2000), LOOPY (Xiang et al. 2002), RAPPER (de Bakker et al. 2003) et PLOP (Zhu et al. 2006). Toutes ces méthodes prennent en compte l'ensemble des atomes. Ainsi, la sélection de la conformation la plus favorable prend en compte la structure et le positionnement des chaines latérales dans le reste de la protéine.

La méthode de prédiction des boucles incluse par défaut dans *Modeller* échantillonne les boucles dans un espace cartésien et cherche à optimiser la fonction d'énergie grâce à une approche couplant gradients conjugués et dynamique moléculaire avec du recuit simulé. La sélection de la conformation associée à la plus faible énergie intervient après 50 à 500 optimisations indépendantes (Fiser et al. 2000).

LOOPY utilise un filtre d'énergie favorisant les conformations ayant un grand nombre de conformations voisines dans l'espace conformationnel. Pour la prédiction d'une boucle, 2000 conformations sont générées pour le squelette polypeptidique, puis filtrées (Xiang et al. 2002).

De même, RAPPER génère 1000 conformations du squelette polypeptidique en se basant sur des tables d'angles (Φ,Ψ). La meilleure candidate est ensuite sélectionnée grâce à une combinaison du champ de force AMBER et d'un modèle de solvatation GBSA (*Generalized Born/Surface Area*) (de Bakker et al. 2003).

PLOP réalise un échantillonnage conformationnel extensif en coordonnées internes (Φ,Ψ) et évalue l'énergie des modèles avec un champ de force OPLS tout-atome couplé à un modèle de solvatation de Born Généralisé et de nouveaux termes dédié à la modélisation de l'hydrophobicité (Zhu et al. 2006).

Enfin, *LoopBuilder* est basé sur l'étape d'échantillonnage de LOOPY et sélectionne la meilleure conformation en utilisant un potentiel de force moyen calculé avec DFIRE puis réalise une minimisation (Zhou and Zhou 2002).

Notre propos n'est pas d'entrer en compétition avec ces méthodes sophistiquées utilisant des champs de force, des techniques de minimisation et des fonctions d'énergie. Notre objectif est uniquement de montrer que notre approche est en mesure de proposer des points de départ structuraux pertinents pour une analyse plus poussée ou un algorithme plus sophistiqué.

Il est important de souligner qu'à nouveau, ce travail de comparaison n'est pas trivial pour deux raisons principales : *(i)* la définition des régions de boucles et *(ii)* l'évaluation de la précision des prédictions. En effet, les structures de connexion associées aux PSLs peuvent correspondre à des boucles plus courtes que 11 résidus mais peuvent aussi faire partie de boucles plus longues. De plus, l'évaluation de notre stratégie est réalisée sur toutes les boucles protéiques de notre jeu de données, soit 24 856 fragments. Inversement, la plupart des méthodes de prédiction des boucles sont évaluées sur des jeux de boucles soigneusement sélectionnés. Par exemple, l'évaluation de *LoopBuider* ne prend pas en compte les structures cristallisées à des pH non standards et les boucles impliquées dans des interactions avec des ligands. Les jeux de boucles peuvent donc être très petits, *e.g.,* l'échantillon de validation de *LoopBuider* pour les boucles de 11 résidus comprends 54 fragments. En conséquence, nos résultats peuvent être affectés par des artefacts dus à la présence de ligands ou à des conditions expérimentales particulières.

Comme exposé précédemment (paragraphe 3.3.4.1), l'évaluation de notre approche repose notamment sur le calcul du Cα RMSD après superposition optimale entre le PSL prédit et la structure locale réelle. Ce type de RMSD est dit *local*. Classiquement, le critère utilisé pour la prédiction des boucles est le RMSD dit *global* calculé en tenant compte de tous les atomes lourds du squelette polypeptidique (Cα, C', O, N) après superposition des extrémités de la boucle prédite et réelle (Fiser et al. 2000). Les analyses de Fiser et collaborateurs permettent d'avoir une idée de la relation entre les deux mesures : le RMSD global est environ égal à 1,5 fois le RMSD local sur les atomes lourds ; par ailleurs, ce dernier est similaire au Cα RMSD local. Dans le but de donner une échelle de comparaison, nous utiliserons donc ce facteur dans la suite de l'analyse.

Pour des boucles de 11 résidus de long, après échantillonnage et sélection du meilleur candidat structural, la méthode de modélisation de *Modeller* atteint un RMSD global moyen de 5,5 Å, LOOPY 3,5 Å, RAPPER 4,9 Å, PLOP 1,0 Å et *LoopBuilder* 2,5 Å. En considérant les 5 candidats structuraux proposés par notre stratégie *SVM_PSSM*, un Cα RMSD local moyen de 3,4 Å est obtenu (Tableau 10, ligne intitulée *secondary structure category C*), soit un RMSD global proche de 5,1 Å en considérant le facteur multiplicateur de 1,5. Parmi ces cinq candidats, un au moins fournit une approximation d'une précision égale à 2,5 Å en moyenne (Cα RMSD local) correspondant à un RMSD global de 3,7 Å (avec le facteur de 1,5). De plus, si nous nous intéressons à présent aux positions bien prédites selon notre critère géométrique (49,5 % des fragments), le RMSD local moyen sur les 5 candidats est de 2,9 Å

(4,4 Å avec le facteur 1,5) et le meilleur candidat fournit en moyenne une approximation de 1,75 Å (2,6 Å avec le facteur 1,5). Ce dernier résultat est d'autant plus intéressant que nous avons développé un indice de confiance permettant d'évaluer directement la qualité de la prédiction et donc d'identifier les fragments bien prédits. Le développement de cet indice et les résultats associés seront présentés dans le paragraphe 4.2.

En conclusion, nos résultats sont comparables à ceux de *Modeller*, LOOPY et RAPPER. En revanche, *LoopBuilder* et PLOP obtiennent de meilleurs modèles de boucles. Néanmoins, les temps de calculs nécessaires à ces algorithmes sont très importants. Le temps CPU moyen pour la modélisation d'une boucle de 11 résidus avec PLOP est d'environ 12 jours avec un processeur (Zhu et al. 2006). *LoopBuilder*, qui a été conçu pour une utilisation optimisée des ressources de calculs, nécessite également plusieurs heures pour construire un modèle (Soto et al. 2008). Ceci met à nouveau en lumière la pertinence de nos résultats : notre approche donne une prédiction instantanée pour une fenêtre de séquence cible. De plus, il est important de souligner que nous ne bénéficions d'aucune contrainte concernant les extrémités des boucles prédites, ni d'aucune information sur le reste de la protéine et la position des chaînes latérales. De plus, nous n'utilisons aucun filtre reposant sur un calcul d'énergie. En conséquence, notre méthode de prédiction des structures locales est compétitive avec les méthodes de pointe dédiées à la prédiction des boucles. Plus important encore, nos prédictions semblent tout à fait en mesure de fournir des informations clés pour enrichir ce champ de recherche. Elles pourraient être utilisées pour filtrer ou orienter les étapes d'échantillonnage des méthodes actuellement les plus performantes.

4.1.5 Conclusion

J'ai présenté ici le développement d'une nouvelle stratégie pour améliorer la prédiction des structures locales protéiques. Nous avons couplé une méthode d'apprentissage sophistiquée, les SVMs, avec un enrichissement des séquences cibles par des informations évolutionnaires obtenues grâce à PSI-BLAST. Pour une fenêtre de séquence cible donnée, nous proposons les cinq candidats structuraux les plus compatibles. Cette approche réduit ainsi fortement la combinatoire des conformations possibles en chaque position d'une séquence protéique. Cette stratégie nous a permis d'obtenir un taux de prédiction de 63,1 % pour 120 classes structurales caractérisant des fragments de 11 résidus de long (évaluation selon un critère géométrique avec un seuil à 2,5 Å). Ce résultat correspond à un gain significatif par rapport à

la méthode de prédiction originale développée par Benros et collaborateurs (Benros et al. 2006) : plus de 11 % en considérant un nombre variable de candidats (4,2 en moyenne) et toujours plus de 7 % en considérant un nombre fixe de 5 candidats (*LR_seq*). De plus, cette amélioration concerne toutes les catégories de PSLs. En effet, le taux de prédiction de 88,3 % des classes structurales est amélioré.
Par ailleurs, nous avons réalisé différentes comparaisons entre notre stratégie et des méthodes de prédiction des structures locales de pointe. L'ensemble de ces comparaisons, bien que non directes, contribue à mettre en avant les capacités prédictives de notre approche et renforce son intérêt dans le cadre de la modélisation de structures protéiques par homologie ou *ab initio*.

En 2004, Pei et Grishin avaient suggéré qu'une description des séquences combinant *(i)* des données évolutionnaires et *(ii)* les préférences structurales des acides aminés, serait une piste prometteuse pour améliorer la prédiction des structures locales. Notre approche réalise ce couplage de façon implicite. De plus, un point important de notre stratégie est que les séquences homologues et le calcul des PSSMs sont réalisés sur la séquence protéique cible entière. Ce choix permet de prendre en compte l'environnement structural et physico-chimique des fragments de séquence et d'apprendre les préférences en acides aminés dans les différentes familles de séquences. Il serait intéressant de comparer ces PSSMs à des PSSMs calculés uniquement à partir des fenêtres de séquence et d'observer si ces derniers peuvent également capturer des interactions et des propriétés à longues distances. L'influence de la qualité des PSSMs serait aussi un point à approfondir. L'analyse de l'impact des alignements non pertinents potentiellement réalisés par PSI-BLAST pourrait permettre d'améliorer encore les performances de prédiction. Les développeurs de PSI-BLAST, Altschul et collaborateurs, ont notamment suggéré qu'un réalignement multiple optimal des séquences avant de recalculer le PSSM pourrait permettre de raffiner ce dernier (Altschul et al. 1997). Enfin, un dernier point délicat est la procédure à suivre dans le cas de séquences cibles avec très peu d'homologues connus. Là encore, une analyse des alignements réalisés par PSI-BLAST pourrait permettre d'identifier ces cas et d'adapter la stratégie de prédiction.

Enfin, il convient de souligner une nouvelle fois la longueur importante des structures locales pour lesquelles ce travail a été réalisé. La plupart des librairies décrivent des fragments de 4 à 9 résidus de long (cf. paragraphe 3.1). La librairie des *I-sites* de Bystroff et Baker caractérisaient des fragments de 3 à 15 résidus. Mais, la méthode de prédiction associée n'a

été évaluée que pour des fragments de 8 résidus (Bystroff and Baker 1998). Récemment, Baeten et associés ont construits une bibliothèque, nommée *BriX*, de 1000 fragments de 4 à 14 résidus de long (Baeten et al. 2008). De même, Sawada et Honda ont développé une base de données, *ProSeg*, de groupes de fragments structuraux de 5, 9, 11 et 15 résidus (Sawada and Honda 2009). Mais pour ces deux derniers exemples, aucune méthode de prédiction n'a été développée. La prédiction de structures locales longues de 11 résidus est donc un travail précurseur.

4.2 Définition d'un indice de confiance

4.2.1 Objectif

Une fois la prédiction réalisée, il est particulièrement utile d'être en mesure d'en estimer sa pertinence. Certaines régions des structures protéiques sont plus difficiles à prédire que d'autres. Un indice de confiance permettant d'identifier ces zones est donc un atout supplémentaire essentiel. Il permet notamment d'imposer de fortes contraintes structurales au niveau des régions prédites avec certitude et de reconnaître les régions nécessitant un raffinement des prédictions (de Brevern et al. 2007).

Le développement d'un indice de confiance permettant de quantifier le degré de prédictibilité structurale d'une séquence a déjà été réalisé dans le cadre de la prédiction des BPs (N_{eq}, paragraphe 3.2.4) (de Brevern et al. 2000; Etchebest et al. 2005) et par Benros *et al.* pour la prédiction des PSLs (paragraphe 3.3.4.1) (Benros et al. 2006). Ces indices reflètent la quantité d'information contenue dans une séquence cible en fonction de la capacité des experts à fournir une prédiction correcte. Je décrirai ici notre stratégie pour le développement d'un indice adapté à la méthode de prédiction *SVM_PSSM*.

4.2.2 Méthode

L'indice de confiance (IC) a été déterminé sur la base des scores calculés par les 120 experts SVMs pour une fenêtre de séquence cible donnée.

Dans ce but, une prédiction des structures locales a été effectuée sur tous les fragments de l'échantillon 2 avec la stratégie *SVM_PSSM*. Pour chaque fragment de séquence, les scores calculés par les experts sont extraits et associés à une prédiction correcte ou une prédiction incorrecte en fonction du critère géométrique à 2,5 Å (cf. Figure 54). Ces données sont utilisées pour entrainer un nouveau SVM (nommé SVM_{IC}) à reconnaître un ensemble de scores menant à une prédiction correcte d'un autre ensemble menant à une prédiction incorrecte. La procédure d'apprentissage utilisée est très similaire à celle réalisée pour

l'entrainement des experts pour la prédiction des structures locales (paragraphe 4.3.1.3.2). Un noyau RBF a été choisi et les paramètres C et γ ont été optimisés par grille de validation croisée.

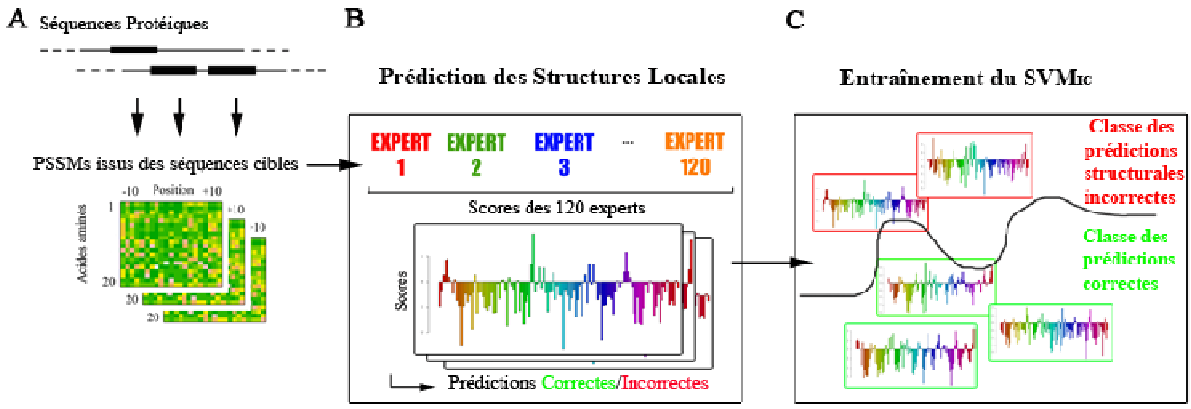

Figure 54. Définition de l'indice de confiance.

A – Conformément à la stratégie *SVM_PSSM* déjà explicitée précédemment, les fenêtres de séquences cibles sont enrichies grâce à la recherche de séquences homologues et représentées par des PSSMs. B – Pour chacune des séquences cibles, les 120 experts calculent chacun un score de décision. Ainsi, pour un échantillon de *N* fragments, *N* profils de 120 scores sont obtenus. En fonction du critère géométrique à 2,5 Å, les profils correspondant à des prédictions correctes sont dissociés des profils associés à des prédictions incorrectes. C – Enfin, un SVM (SVM_{IC}) est entrainé pour reconnaître ces deux types de profils.

Finalement, pour une prédiction donnée, l'IC est directement défini comme étant la valeur de décision v_s ou *score* calculé par le SVM_{IC}. Ce score quantifie la distance à l'hyperplan séparant les bonnes prédictions des mauvaises. Ainsi, un IC très positif caractérise une bonne confiance dans la prédiction. En revanche, un IC très négatif doit être associé à des prédictions difficiles. Les résultats obtenus avec l'échantillon 1 ont été utilisés pour modéliser la relation entre l'IC et le taux de prédiction. Cette relation a été vérifiée en utilisant l'échantillon 3 des protéines.

4.2.3 Résultats

Les IC obtenus sur l'échantillon 1 grâce au modèle SVM_{IC} ont été divisés en 30 zones équivalentes. Les 8 premières et les 5 dernières catégories représentant seulement 1,15 et 0,95 % des prédictions respectivement ont ensuite été rassemblées. Finalement, 19 catégories d'IC ont été définies.

Les taux de prédiction des structures locales obtenus sur les échantillons 1 et 3 sont représentés en fonction des 19 catégories d'IC en Figure 55A. Les catégories sont ordonnées

d'un niveau de confiance faible à un niveau élevé. La courbe caractérisant la relation entre les catégories d'IC et les taux de prédiction est clairement sigmoïde. Quelque soit l'échantillon considéré, le taux de prédiction globale est élevé (61,3 % et 63,1% respectivement) et la distribution des taux en fonction des catégories d'IC est presque identique. Les IC inférieurs à -0,58 (catégories 1 à 5) sont associés à une qualité de prédiction médiocre, avec des taux de prédiction allant de 20 à 30 %. A l'inverse, les IC supérieurs à 1,10 (catégories 15 à 19) sont associés à des prédictions presque parfaites, avec des taux allant de 83,9 à 96,8 %. Entre ces deux extrêmes, le taux de prédiction augmente rapidement en fonction des catégories d'IC. Ainsi, nous montrons ici que notre modèle SVM_{IC} permet d'obtenir un IC lié à la fiabilité des prédictions : plus la valeur de l'IC est élevée, plus la prédiction est fiable. En conséquence, l'IC proposé est un outil pertinent pour l'évaluation directe de la qualité de la prédiction.

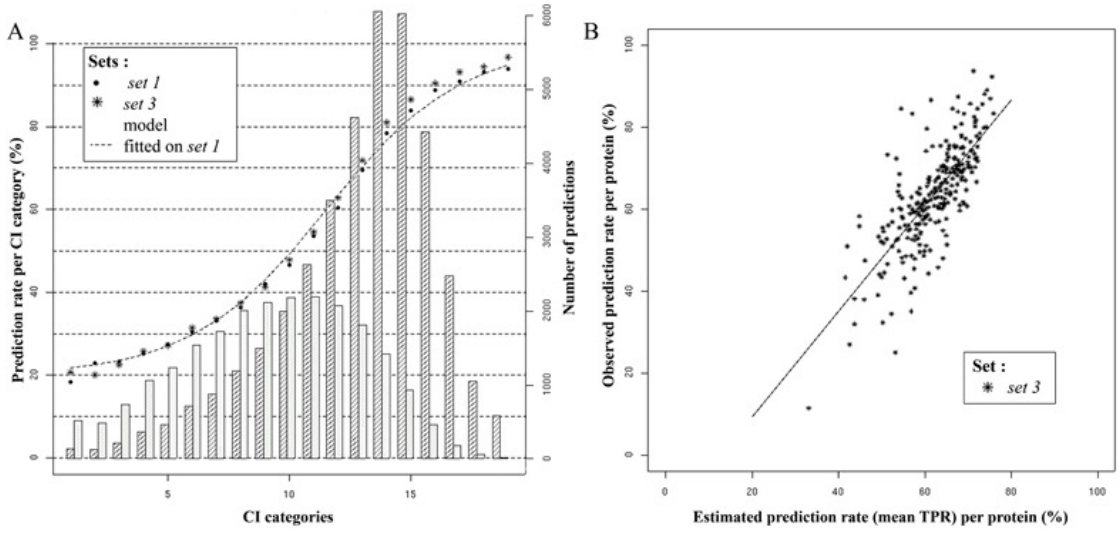

Figure 55. Validation de l'indice de confiance.

A: Représentation du taux de prédiction associé à chaque catégorie d'IC sur les échantillons 1 et 3. Un modèle permettant de caractériser la relation entre taux de prédiction et catégories d'IC a été établi sur l'échantillon 1. Il est présenté en courbe pointillée. De plus, un histogramme montre le nombre de prédictions correctes (hachures foncées) et le nombre de prédictions incorrectes (hachures claires) pour chaque catégorie d'IC sur l'échantillon 3. Les catégories d'IC sont numérotées de 1 à 19 est correspondent aux intervalles suivants :]-∞, -1,32],]-1,32, -1,14],]-1,14, -0,95],]-0,95, -0,77],]-0,77, -0,58],]-0,58, -0,39],]-0,39, -0,21],]-0,21, -0,02],]-0,02, 0,17],]0,17, 0,36],]0,36, 0,54],]0,54, 0,73],]0,73, 0,92],]0,92, 1,10],]1,10, 1,29],]1,29, 1,48],]1,48, 1,67],]1,67, 1,85],]1,85, +∞]. B : Représentation du taux de prédiction observé par protéine de l'échantillon 3 en fonction du Taux de Prédiction Théorique (TPT) estimé moyen. Le TPT est calculé à partir de la simple connaissance de l'IC. La relation linéaire entre ces 2 variables est montrée par la ligne noire ($r_{Pearson}$=0,77, p-value < $2{,}2\ 10^{-16}$). Figure extraite de (Bornot et al. 2009).

La relation sigmoïde entre les catégories d'IC et le taux de prédiction peut être caractérisée en effectuant une régression linéaire après transformation logarithmique des résultats de l'échantillon 1. A partir de cette relation, un Taux de Prédiction Théorique peut être calculé (TPT) pour une prédiction donnée en connaissant la catégorie de l'IC :

$$\mathrm{TPT} = \frac{4}{5}\left(\frac{1}{1+e^{-0.36(CI\ \mathrm{category}-11.53)}} + \frac{1}{4} \right)$$

Ce modèle explique 96 % des données de l'échantillon 1 (le coefficient de détermination est égal à 0,96) et la distribution des résidus de la régression suivent une loi normale (Lilliefors test; R software (Ihaka and Gentleman 1996)). De plus, ce modèle construit à partir de l'échantillon 1 a été validé sur l'échantillon 3. Le TPT moyen calculé sur toutes les prédictions de l'échantillon 3 est de 61,8 %. Cette valeur est très poche de la valeur réelle observée, *i.e.,* 63,1 %. La Figure 55B présente la correspondance entre le TPT estimé moyen pour chaque protéine de l'échantillon 3 et le taux de prédiction observé. Une relation linéaire statistiquement significative est observée ($r_{Pearson}$=0,77, *p-value* < 2,2 10^{-16}). De même, une corrélation significative est observée entre le TPT estimé moyen pour chaque classe structurale et le taux de prédiction ($r_{Pearson}$=0,86, *p-value* < 2,2 10^{-16}). Ainsi, l'indice de confiance défini dans ce travail couplé au modèle d'estimation du taux de prédiction théorique permet une évaluation directe de la qualité d'une prédiction sur une échelle allant de 0 à 100%.

4.2.4 Conclusion

Pour chaque prédiction, l'indice de confiance défini dans ce travail et couplé au modèle d'estimation du taux de prédiction théorique permet une évaluation directe de la qualité des propositions structurales sur une échelle allant de 0 à 100% (voir Figure 56).

Les indices de confiance sont des outils particulièrement utiles dans le cadre d'une prédiction par homologie incluant des prédictions de structures locales, mais aussi pour des constructions *de novo* de modèles protéiques. Une procédure hiérarchique peut être envisagée dans laquelle, dans un premier temps, les régions considérée comme prédites avec précision sont sélectionnées et fixées, puis dans un second temps les régions de plus faible informativité

seront examinées plus en détail en prenant en compte un plus grand nombre de candidats par exemple. Ce type de procédure permettra une réduction significative de l'espace conformationnel global à échantillonner pour une protéine donnée.

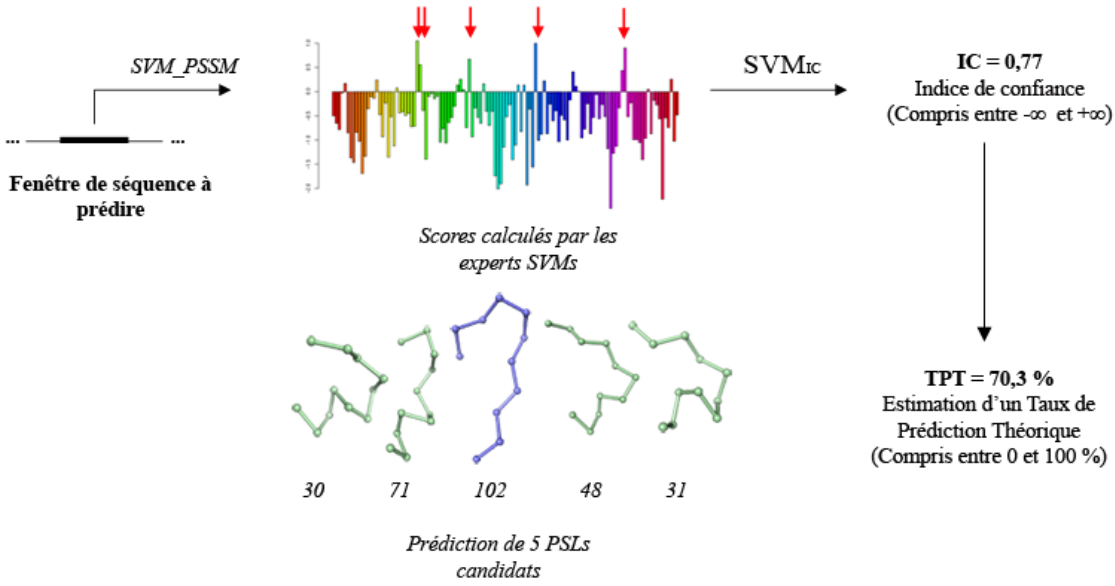

Figure 56. Exemple d'une prédiction structurale accompagnée d'une estimation de sa fiabilité.

Pour une fenêtre de séquence cible, la méthode de prédiction des structures locales SVM_PSSM calcule des scores d'adéquation entre la séquence et chacune des 120 classes structurales. Les 5 classes les plus compatibles sont alors proposées : dans cet exemple, les PSLs 30, 71, 102, 48 et 31. Le modèle SVM_{IC} utilise ensuite les 120 scores de compatibilité pour calculer un IC. Ici, l'IC égale 0,77. Finalement, grâce à la relation observée entre le taux de prédiction et les catégories d'IC, ce dernier permet d'estimer un TPT. Ici, l'IC de 0,77 appartient à la catégorie n°13 (voir la légende de la Figure 55), soit un TPT de 70,3%. D'après ce TPT, notre fenêtre de séquence fait donc partie des zones relativement faciles à prédire, généralement associées à des taux de prédictions correctes de l'ordre de 70%.

5. Flexibilité des structures protéiques

La vision des structures protéiques figées est si prévalente qu'il est aisé d'oublier que les protéines ne sont pas des macromolécules rigides.
Les protéines sont flexibles et leur capacité à se déformer est souvent essentielle à la réalisation de leurs fonctions. Cette flexibilité a pu être observée dès l'obtention des premières structures cristallographiques : en 1964, Perutz et collaborateurs, travaillant sur la cristallisation de l'hémoglobine, observèrent une légère modification de la molécule lorsque celle-ci était liée ou non à l'oxygène. Cette étude du changement de conformation de l'hémoglobine en présence d'un ligand fut la première de la sorte (Levinthal 1966).

Je présenterai tout d'abord les notions de flexibilité et de désordre. Leur importance fonctionnelle ainsi que les mécanismes impliqués dans leur régulation seront ensuite abordés. Enfin, je présenterai les techniques bioinformatiques actuelles dédiées à l'analyse et la prédiction de la flexibilité.

5.1 Notions de flexibilité et de désordre

Comme nous l'avons vu dans la section 2, la vision classique des protéines est une séquence en acides aminés se repliant en une structure 3D optimale et fonctionnelle. Cette structure est dite *ordonnée*. Cette vision statique est notamment propagée par les modèles cristallographiques. Aujourd'hui largement utilisés par la communauté scientifique, ils présentent uniquement l'ensemble de coordonnées atomiques correspondant le mieux aux données expérimentales (Huang and Montelione 2005). Or d'autres techniques expérimentales, comme la RMN, ont fait émerger une nouvelle vision : les protéines peuvent être vues comme un ensemble de différentes conformations à l'équilibre (James and Tawfik 2003; Lange et al. 2008). Ces changements de conformations au sein des structures se caractérisent non seulement par des mouvements de boucles et de chaînes latérales en surface mais également par des mouvements collectifs de régions entières au cœur des protéines. Ainsi, bien que la connaissance des structures reste un outil essentiel pour une meilleure compréhension du fonctionnement des protéines, il est nécessaire d'enrichir cette information

par des données relatives à leur flexibilité et leurs propriétés dynamiques (Lavery and Sacquin-Mora 2007).

Par ailleurs, durant ces dix dernières années, des protéines extrêmement flexibles ont été largement étudiées : les protéines dites *désordonnées*. Ces protéines n'ont pas de structures 3D. Elles existent en tant qu'ensemble dynamique au sein duquel les positions des atomes et les angles dièdres varient de façon significative au cours du temps sans état d'équilibre spécifique (Vucetic et al. 2005). Le terme de désordre décrit une réalité complexe allant du *random coil*, associé à des modifications de conformations très rapide, au *globule fondu* (*molten globule* en anglais) présentant des structures secondaires développées mais pas ou peu d'interactions tertiaires. Par ailleurs, une protéine peut être entièrement désordonnées ou seulement contenir des régions désordonnées plus ou moins longues (Receveur-Brechot et al. 2006). Le désordre est présent au sein de tous les règnes (Ward et al. 2004; Chen et al. 2006). La fréquence des segments désordonnés de plus de 30 résidus serait de plus de 33 % chez les eucaryotes. Cette fréquence serait encore plus importante chez les virus, mais moins élevée chez les bactéries.

5.2 Importance fonctionnelle de la flexibilité structurale des protéines

La flexibilité des protéines *ordonnées* est essentielle pour la réalisation de leurs fonctions biologiques et la vie de la cellule. Leurs propriétés dynamiques sont au cœur des phénomènes d'interaction et de reconnaissance moléculaire (cf. Figure 57) (Peng et al. 2007; Boehr and Wright 2008; Dunker et al. 2008). Depuis plus de 50 ans, le modèle d'*Ajustement Induit* supporte l'hypothèse selon laquelle l'interaction initiale entre une protéine et son partenaire peut induire un changement conformationnel (Koshland 1958). Depuis, plusieurs années, un autre modèle gagne de plus en plus de crédit, le modèle de *Sélection Conformationnelle.* Selon ce modèle, la protéine non-liée existe en tant qu'ensemble conformationnel dans un équilibre dynamique. L'arrivée d'un ligand peut favoriser la stabilisation d'une conformation peu peuplée et de plus haute-énergie, et ainsi modifier l'équilibre (James and Tawfik 2003; Boehr and Wright 2008; Lange et al. 2008).

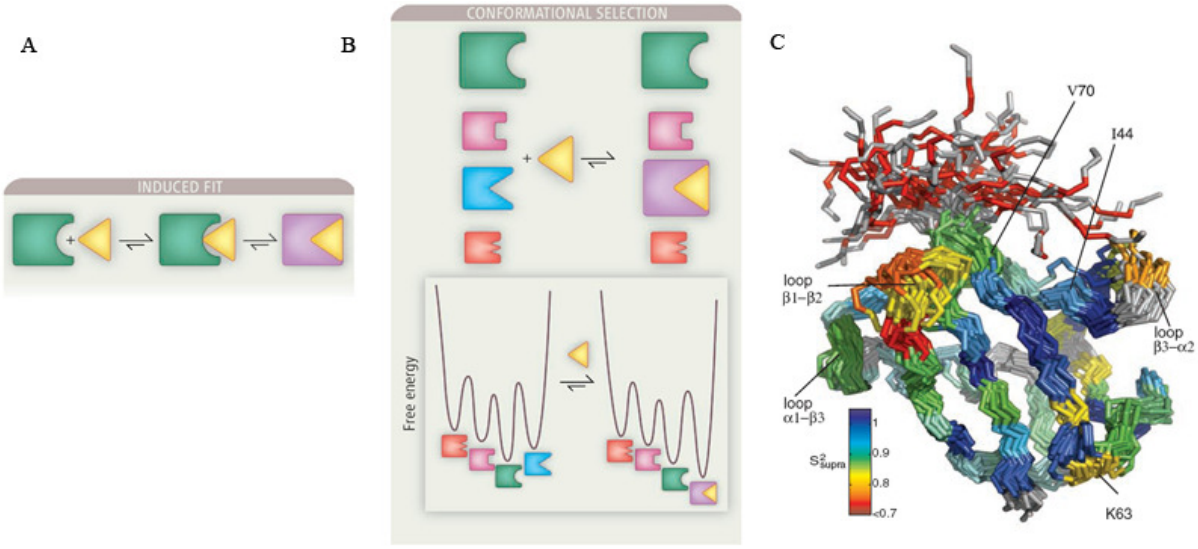

Figure 57. La flexibilité des protéines joue un grand rôle dans les mécanismes de reconnaissance moléculaire.

Deux visions des mécanismes de reconnaissance moléculaire sont présentées. A – L'Ajustement Induit (ou *Induced Fit* en anglais), B et C – La Sélection Conformationnelle (*Conformational Selection* en anglais). Voir le texte ci-dessus pour les explications. C – Ensemble structural de l'Ubiquitine résolu par RMN. Seul le squelette polypeptidique est représenté. Les couleurs illustrent le degré de mobilité des résidus. L'ensemble structural recouvre complètement l'hétérogénéité observée dans 46 structures cristallographiques, la plupart caractérisant des formes complexées. Le modèle de la Sélection Conformationnelle suffit donc à expliquer la dynamique impliquée dans le mécanisme de reconnaissance moléculaire de l'Ubiquitine. Les figures A et B sont extraites de (Boehr and Wright 2008), la figure C de (Lange et al. 2008).

Par exemple, les propriétés dynamiques des enzymes sont essentielles au processus de catalyse. Elles sont impliquées non seulement dans la reconnaissance et la liaison des substrats mais également dans la vitesse de réaction et la libération des produits (Eisenmesser et al. 2005; Boehr et al. 2006a; Boehr et al. 2006b). La flexibilité des protéines est aussi impliquée dans leur stabilité ou encore dans les phénomènes d'agrégation au cœur des maladies de Parkinson et d'Alzheimer et des désordres liés aux prions (Dobson 2003).

Certaines protéines *désordonnées* remplissent leur fonction sans structuration tertiaire (*e.g.,* chaines entropiques (Tompa 2002)). Toutefois, la majorité se structure en se liant à une (des) protéine(s) partenaire(s) (Receveur-Brechot et al. 2006). Dans ce cadre, leur capacité à lier spécifiquement différentes cibles, la possibilité de régulation précise de la liaison ainsi que leur aptitude à former de grandes interfaces intermoléculaires tout en conservant des séquences relativement courtes et donc en limitant la place prise dans la cellule et dans le génome, sont des avantages importants (Dyson and Wright 2005; Fink 2005). Les fonctions majeures associées aux régions désordonnées impliquent tout particulièrement la liaison à l'ADN pour faciliter des processus tels que la transcription ou la réparation. D'autres

processus comme la régulation, la signalisation, le cycle cellulaire ou encore le développement semblent liés aux propriétés de désordre des protéines (Ward et al. 2004).

5.3 Caractérisation de la flexibilité

Ainsi, les protéines sont des objets dynamiques. Les mouvements au sein de leurs structures couvrent un large spectre d'énergie (0,1 à 100 kcal/mol-1), de temps (10^{-15} / 10^{4}s) et d'amplitudes (0,1 à 100 Å voire plus) (cf. Figure 58). De plus, ils sont interdépendants et couplés. La caractérisation de la flexibilité des protéines n'est donc pas triviale.

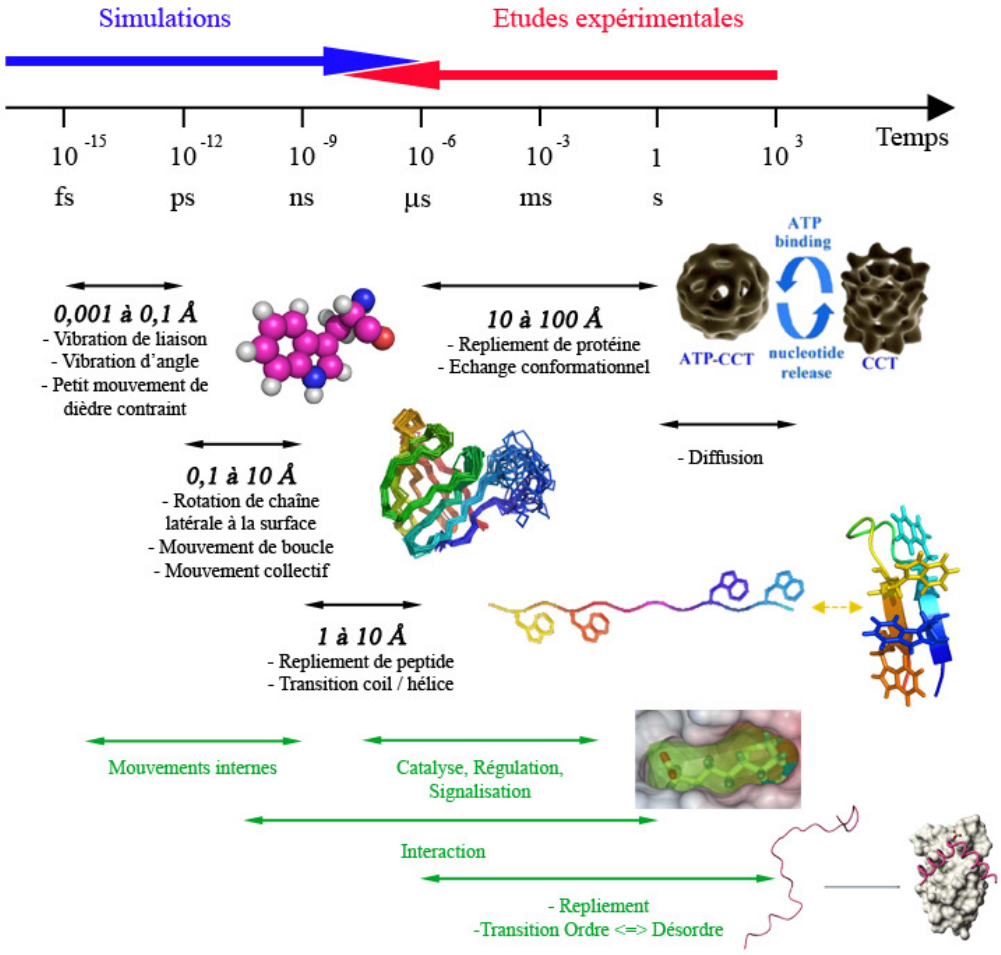

Figure 58. Les mouvements au sein des protéines couvrent un large spectre de temps et d'amplitudes.

Cette figure a été réalisée d'après le cours de Patrick Fuchs sur la « Dynamique Moléculaire » donné dans le cadre du M1 SPGF de L'Université Denis Diderot - Paris 7 et d'après le cours de Carine Van Heijenoort sur « La RMN et les mouvements moléculaires » donné dans le cadre du M2 de Biophysique de l'Université Pierre et Marie Curie. La figure du CCT (en haut à droite) est extraite de (Llorca et al. 2001). Le peptide riche en W déplié est tiré de http://www.lofar.org/BlueGene/Suits.pdf. Le ligand dans son site de liaison est extrait de (Doppelt-Azeroual 2009). L'illustration du repliement induit d'une protéine désordonnée est extraite de (Dyson and Wright 2005). Les autres figures ont été réalisées avec pymol (DeLano 2002).

5.3.1 Différentes visions selon l'échelle de temps considérée

En fonction de l'échelle de temps considérée, différents types de mouvements peuvent être observés (Boehr et al. 2006a) (cf. Figure 58).

De la picoseconde à la nanoseconde, des mouvements rapides du squelette polypeptidique et des chaînes latérales ont lieu. De la microseconde à la milliseconde, des mouvements de plus grande amplitude sont observés comme le repliement de protéines, des échanges conformationnels impliqués dans la liaison à des substrats ou encore dans le phénomène de catalyse. Il n'existe pas de relation directe entre les mouvements ayant lieu à l'échelle de la picoseconde-nanoseconde et ceux ayant lieu de la microseconde à la milliseconde. Des expériences sur des enzymes ont montré que la liaison d'un ligand pouvait influencer (ou non) les mouvements à l'échelle de la *ps-ns* tout en modifiant ou non les mouvements à l'échelle de la *µs-ms* (Freedberg et al. 2002; Butterwick et al. 2004; Boehr et al. 2006a). La Figure 59 présente une étude de la Ribonucléase HI (RNase HI) de *Thermus thermophilus*. Les résidus les plus flexibles considérant les mouvements rapides ne sont par forcément les plus flexibles aux temps plus lents.

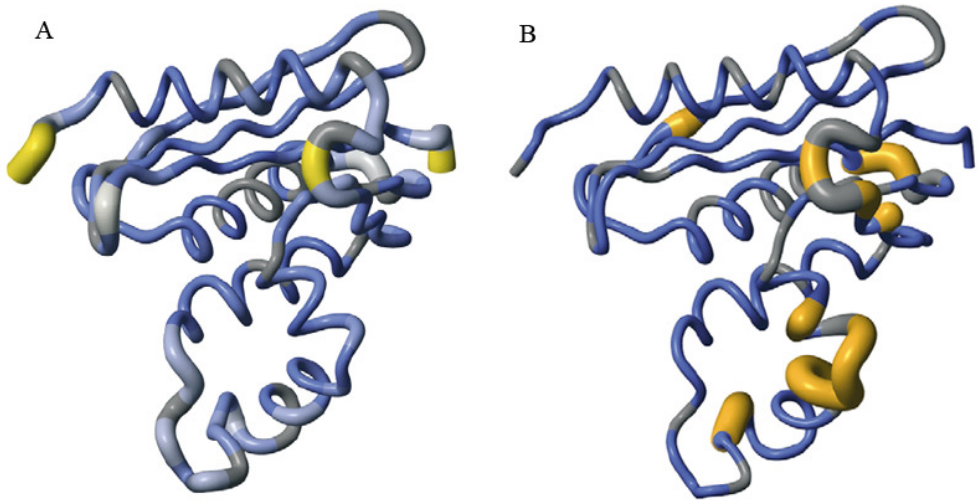

Figure 59. Etude de la flexibilité de la RNase HI de Thermus thermophilus à différentes échelles de temps.

A – Propriétés dynamiques de la RNase HI à l'échelle de la *ps-ns*. La flexibilité est mesurée grâce aux paramètres d'ordre s^2, plus le s^2 est petit plus le résidu est considéré flexible (voir paragraphe 5.3.3.2). Le bleu foncé correspond à un s^2 proche de 1, le bleu clair à un s^2 proche de 0,5 et le jaune à un $s^2 <$ 0,5. B – Propriétés dynamiques à l'échelle de la *µs – ms* à. Les résidus affectés par des échanges conformationnels sont en orange, les autres en bleu. Le gris correspond à des résidus pour lesquels les données étaient insuffisantes. Figure extraite de (Butterwick et al. 2004).

5.3.2 Différentes visions selon nombre de résidus considérés

De même, à l'échelle d'un résidu, la flexibilité est classiquement associée à la *mobilité* de ce résidu au sein de la structure protéique. En revanche, à l'échelle de fragment de structure ou de chaînes polypeptidiques plus longues, les mouvements corrélés des résidus doivent être pris en compte et le concept de *déformabilité* émerge (voir Figure 60) (Kovacs et al. 2004). Les concepts de mobilité et de déformabilité sont complémentaires mais ils ne sont pas nécessairement associés. Par exemple, une région charnière peut subir des modifications conformationnelles (*déformation*) sans fluctuations (*mobilité*) tandis qu'une région présentant des mouvements de *corps rigide* peut montrer des fluctuations de large amplitude sans subir aucune déformation.

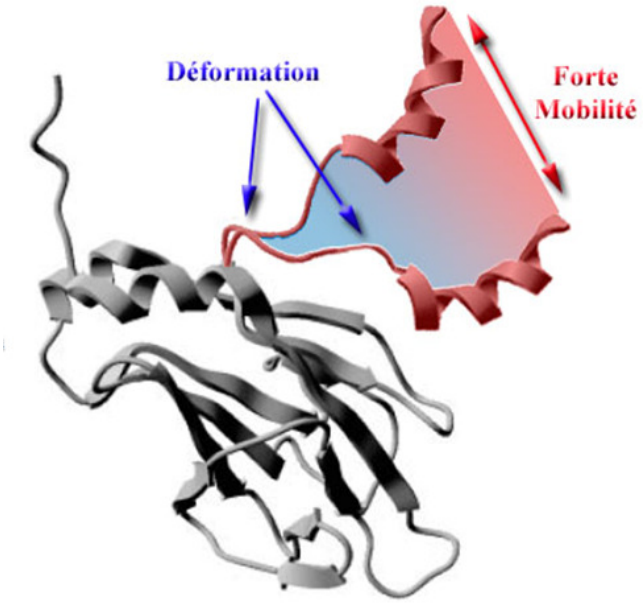

Figure 60. Différence entre mobilité et déformation.

Un cas extrême est présenté pour illustrer la différence entre mobilité et déformation. L'hélice colorée en rouge présente une très forte mobilité mais ne se déforme pas. A l'inverse, la boucle liant l'hélice au reste de la structure se déforme largement mais ne présente pas de mouvement de large amplitude.

5.3.3 Caractérisation expérimentale

Deux principales sources d'informations expérimentales nous renseignent sur la flexibilité des protéines : les facteurs de température cristallographique et la RMN.

5.3.3.1 Les facteurs de températures cristallographiques

Lors de la résolution des structures protéiques par cristallographie, des facteurs de température (appelés aussi *B-facteurs* ou *Debye-Waller factors*) sont calculés pour chaque atome. Ces B-facteurs représentent l'incertitude quant à la position des atomes (Clore and Schwieters 2006). Ainsi, ils reflètent les vibrations thermiques et le désordre statique au sein du cristal. Ils fournissent donc une mesure de la *mobilité* des atomes au sein des protéines. Il convient néanmoins de ne pas oublier que cette mesure peut être entachée de bruit : les B-facteurs dépendent en partie de la résolution de la structure et des procédures de raffinement (Linding et al. 2003; Schlessinger and Rost 2005; Hinsen 2008). Pour comparer les structures entre elles, il est donc nécessaire de normaliser ces valeurs. Des biais peuvent, néanmoins, persister. Par exemple, des contacts au sein du cristal ou l'ajout de molécules stabilisantes permettant la cristallisation peuvent notamment rigidifier une région naturellement très flexible. Toutefois, les B-facteurs restent très largement utilisés pour les études bioinformatiques de la flexibilité, car très accessibles.

Il faut de plus noter que la cristallographie peut aussi apporter des informations sur les échanges conformationnelles au sein des protéines. En effet, de nombreux cristaux de la PDB caractérisent les mêmes protéines dans des conformations différentes, *e.g.,* liées ou libres (Krebs and Gerstein 2000).

5.3.3.2 Mesures de la flexibilité par Résonance Magnétique Nucléaire (RMN)

La RMN (voir paragraphe 2.3.1) est une technique expérimentale puissante pour l'étude de la dynamique des protéines. Les mouvements observables en RMN couvrent une large gamme de temps de la picoseconde à la seconde et plus (Boehr et al. 2006a). Ainsi, l'étude des vitesses de relaxation permet d'extraire des paramètres d'ordre s^2. L'approche la plus populaire pour les calculer est celle de Lipari et Szabo (Lipari and Szabo 1982). Ces paramètres mesurent la corrélation des mouvements des dipôles N-H ou C-H avec le reste de la protéine. Un s^2 de 1 signifie une forte corrélation et une totale restriction des mouvements. A l'inverse, un s^2 de 0 indique un mouvement isotropique sans restriction. Les paramètres d'ordre sont l'outil le plus classique en RMN pour étudier les mouvements à l'échelle de la *ps-ns*. Cependant, comme le B-facteur, le paramètre d'ordre a aussi des limites. Le calcul de

Lipari et Szabo suppose une diffusion rotationnelle globale lente des protéines en solution. Or, des mouvements internes lents peuvent exister et être difficilement séparables du mouvement global de la protéine, menant ainsi à une surestimation des s^2. De même, des protéines avec des temps de corrélation très différents pour leur diffusion rotationnelle peuvent être difficilement comparables (Idiyatullin et al. 2003). Par ailleurs, ces mesures sont peu mises à disposition de la communauté scientifique. La *Biological Magnetic Resonance Data Bank* (*BMRB*) s'est fixé le but de collecter et disséminer les mesures quantitatives issues de la RMN en collaboration avec la PDB (Markley et al. 2008). Toutefois, à ce jour, les s^2 ne sont disponibles que pour 44 protéines.

L'étude des mouvements correspondant au spectre *µs-ms* nécessite la réalisation d'expériences dites de Relaxation-Dispersion (Eisenmesser et al. 2005; Korzhnev and Kay 2008). Entre ces deux fenêtres, à l'échelle de la µs, la technique du Couplage Dipolaire Résiduel (*Residual Dipolar Coupling* en anglais ou RDC) apporte également des informations (Lakomek et al. 2008).

Il est à noter que cette liste n'est pas exhaustive. D'autres techniques permettent d'étudier la flexibilité des protéines comme des techniques de fluorescence ou encore le suivi d'échanges hydrogène/deutérium par spectrométrie de masse (Boehr et al. 2006a). Néanmoins, à ma connaissance, les données issues de ces mesures ne sont pas mises à disposition de la communauté scientifique.

5.3.4 Analyse *in silico* à partir de la structure 3D

De nombreuses analyses *in silico* permettent d'étudier la flexibilité des structures protéiques. Les plus classiquement utilisées sont la dynamique moléculaire, les modes normaux et l'échantillonnage conformationnel. Je présenterai tout d'abord la technique de la dynamique moléculaire sur laquelle repose une partie de notre analyse de la flexibilité (cf. paragraphe 7) puis j'aborderai brièvement les autres méthodes d'analyse.

5.3.4.1 Simulation de dynamique moléculaire

Au cours des simulations de dynamique moléculaire, les configurations successives du système sont générées en intégrant les lois du mouvement de la mécanique classique newtonienne. Le résultat est une trajectoire décrivant les variations des positions et des vitesses des particules du système au cours du temps (Leach 2001). Chaque atome est représenté par une masse ponctuelle. Son mouvement est déterminé par l'ensemble des forces

exercées sur lui par les autres atomes en fonction du temps. Ces forces sont évaluées grâce à un champ de force empirique décrivant l'énergie potentielle du système. Les paramètres de ce champ, comme les charges atomiques, les valeurs d'équilibre pour les longueurs et angles de liaison, sont ajustés de manière à reproduire les calculs de mécaniques quantiques et des données expérimentales.

La limite principale de la dynamique moléculaire est le temps de simulation envisageable avec des moyens de calculs actuels. Quelques simulations allant jusqu'à la microseconde ont été réalisées mais restent rares (Duan and Kollman 1998; Grossfield et al. 2008). De nos jours, les simulations sont en général de l'ordre de la dizaine de nanoseconde (Okazaki and Takada 2008; Chen 2009). La pertinence de cette technique pour analyser le fonctionnement des protéines en accord avec des résultats expérimentaux a néanmoins été démontrée à plusieurs reprises (Dodson et al. 2008). Elle est de plus utilisée dans les protocoles de résolution de structures protéiques 3D par cristallographie et par RMN, ou encore dans la prédiction de modèle structuraux.

5.3.4.2 D'autres méthodes d'étude de la flexibilité à partir de la structure

La technique des modes normaux est également largement utilisée par la communauté scientifique. Elle permet d'étudier les mouvements collectifs au sein des structures protéiques. Elle repose sur une approximation harmonique de la fonction d'énergie potentielle autour d'un minimum énergétique. Cette approximation permet la résolution analytique des équations du mouvement autour d'une conformation d'énergie minimale (Suhre and Sanejouand 2004; Sanejouand 2007). Abagyan et collaborateurs se sont notamment appuyés sur cette technique pour déduire un indice de *déformabilité* et développer une méthode de prédiction associée, *i.e.,* DFprot (Kovacs et al. 2004; Garzon et al. 2007).

De même, des techniques d'exploration de l'espace conformationnel comme la technique du Monte Carlo ou le Recuit Simulé sont désormais classiques. Elles permettent des études de la flexibilité et de la stabilité des protéines (de Groot et al. 1997; Leach 2001; de Brevern et al. 2005; Fuchs et al. 2006; Doucet and Pelletier 2007). Des techniques d'interpolation entre deux structures cristallographiques ont également été proposées et ont récemment été comparées (Weiss and Levitt 2009). Une banque de données, MolMovDB, propose une classification de mouvements interpolés (Flores et al. 2006).

Il convient aussi de citer l'algorithme FIRST reposant sur la théorie des graphes et une analyse du nombre de degrés de liberté (Jacobs et al. 2001). Il permet d'identifier très rapidement les régions flexibles et rigides de la structure et a récemment été amélioré pour

tenir compte des propriétés dynamiques des liaisons non covalentes (Mamonova et al. 2005). Enfin, certaines méthodes proposent de déterminer la flexibilité d'une structure en appliquant des perturbations locales. Ainsi, Sacquin-Mora et Lavery proposent une mesure de la rigidité des résidus dérivée de la force nécessaire à appliquer pour bouger chacun des résidus par rapport au reste de la structure (Sacquin-Mora and Lavery 2006; Sacquin-Mora et al. 2007). De même, la méthode RIP (*Rotamerically Induced Perturbation*) induit des mouvements à moyenne échelle en perturbant les chaînes latérales des résidus lors de simulations de dynamique moléculaire de l'ordre de la dizaine de picosecondes (Ho and Agard 2009).

5.4 Prédiction de la flexibilité à partir de la séquence

5.4.1 Relation séquence-structure

Radivojac *et al.* ont analysé la composition en acides aminés des régions ayant un B-facteur élevé (résidu avec B-facteur normalisé > 2) et les ont comparé d'une part aux régions avec un faible B-facteur (B-facteur normalisé ≤ 2) et d'autre part aux régions désordonnées (définies comme les régions pour lesquelles aucune coordonnées cristallographique ne peut être obtenue) (Radivojac et al. 2004). La composition des régions avec un fort B-facteur est plus proche de la composition des régions désordonnées sans être totalement similaire. Au sein des régions très flexibles avec de forts B-facteurs et des régions désordonnées, les résidus classiquement assez enfouis comme le tryptophane, la phénylalanine, la tyrosine et l'isoleucine sont sous-représentés. En revanche, la sérine, la proline, la glutamine, le glutamate et la lysine sont sur-représentés. Les régions à forts B-facteurs sont spécifiquement enrichies en glycine, aspartate et asparagine. De manière générale, les régions flexibles souvent plus hydrophiles et de charge nette absolue élevée. Ce biais compositionnel supporte l'hypothèse de la prédictibilité des régions flexibles à partir de la séquence protéique.

5.4.2 Méthodes de prédiction à partir de la séquence

Deux catégories de méthodes de prédiction à partir de la séquence peuvent être décrites : les méthodes dédiées à la prédiction de la *mobilité* des résidus et celles dédiées à la prédiction de la *déformabilité* de fragments de séquences (voir Tableau 12).

Tableau 12. Les méthodes de prédiction de la flexibilité.

	Citation	Nom de la Méthode	Mesure de la Flexibilité	Informations utilisées	Méthode d'apprentissage	Prédiction	Evaluation				
							corrélation	Taux de Prediction (%)	MCC	Acc (%)	Cov (%)
Mobilité	Karplus, 1985		B-facteurs	-séquence	Utilisation directe de la flexibilité moyenne des résidus	B-facteurs théoriques	0,34[a]	Ø	Ø	Ø	Ø
	Vihinen, 1994	VTR	"	-séquence	Utilisation directe de la flexibilité moyenne des résidus en fonction de la rigidité de leurs voisins	B-facteurs théoriques	0,33	Ø	Ø	Ø	Ø
	Radivojac, 2004	PS	"	-Prédiction de SII -Profiles évolutionnaires	Régression Logistique	B-facteurs théoriques	0,43	69,7	Ø	Ø	Ø
	Yuan, 2005		"	-Profiles évolutionnaires	SVR	B-facteurs théoriques + 2 classes	0,53	71,3/69,5[b]	Ø	Ø	Ø
	Schlessinger, 2005	PROFbval	"	-Profiles évolutionnaires -Prédiction de SII -Prédiction de l'accessibilité	Réseau de Neurones	2 classes	0,44/0,50[c]	Ø	Ø	63,1/70,1[d]	46,2/73,9[d]
	Chen, 2007 b		"	-Profiles evolutionnaires -Estimation des vitesses évolutives - Profiles d'hydrophobicité	SVM multi-classes	3 classes	0,57	Ø	Ø	59,2	65,0
	Zhang, 2009	PredRSA9	"	-Prédiction de l'accessibilité	Régression Linéaire	B-facteurs théoriques	0,53/0,55[e]	Ø	Ø	Ø	Ø
	Gu, 2006	Wiggle	Fluctuations à partir d'un Réseau Gaussien	-modèle HMM basé sur des données évolutionnaires	SVM	2 classes	Ø	66,0/76,5[f]	Ø	37,1/49,0[f]	70,5/78,3[f]
	Tartaglia, 2007	CamP	facteurs de protection (échanges $H/^2H$)	-séquence seule	SVM	Facteurs de Protection théoriques	0,76	Ø	Ø	Ø	Ø
Déformabilité	Boden, 2006		Comparaison de structures cristallographiques	-Profiles PSI-BLAST -Prédiction de SII	Réseau de Neurones probabiliste en cascade	2 classes	Ø	Ø	0,13	12,0	76,0
	Chen, 2007 a	FlexRP	"	-Profiles PSI-BLAST -Prédiction de SII - Fréquence des pairs d'acides aminés à une distance *k*	Régression Logistique	2 classes	Ø	79,5	0,51	Ø	Ø

ACC = TP / (TP+FP)
COV = TP/ (TP+FN)

SII:Structures Secondaires

[a] Evaluation qualitative présentée sur un seul exemple dans l'article. Le coefficient de corrélation a été évalué a posteriori par (Vihinen, 1994).
[b] Différents seuils pour les 2 classes de flexibilité: 0,03/-0,23.
[c] Evaluation sur le jeu de structures de Schlessinger/ Evaluation sur le jeu de (Yuan, 2005).
[d] Seuils strict (0,03)/Non Strict (-0,3) entre les 2 classes de flexibilité.
[e] Sur deux jeux de données différents. Le second est celui de (Yuan, 2005).
[f] jeu de données classique / jeu de données de protéines plus petites que 200 résidus.

Les premières définissent le plus souvent la *mobilité* des résidus en termes de B-facteurs mesurés au niveau des Cα lors des expériences de cristallographie (paragraphe 5.3.3.1). Les méthodes pionnières reposaient sur la séquence seule et une analyse des valeurs moyennes de B-facteurs par type d'acide aminé (Karplus and Schulz 1985; Vihinen et al. 1994). Aujourd'hui, la plupart des méthodes utilisent des informations évolutionnaires, la prédiction des structures secondaires ou encore de l'accessibilité au solvant et exploitent des méthodes de régression élaborées comme la régression logistique, la régression à vecteurs supports (SVR) ou encore des réseaux de neurones (Radivojac et al. 2004; Schlessinger and Rost 2005; Yuan et al. 2005; Chen et al. 2007b; Zhang et al. 2009). De façon alternative, la méthode CamP repose sur des valeurs de facteurs de protection obtenues *via* des expériences d'échanges hydrogène/deutérium. Ces facteurs semblent pouvoir mesurer des mouvements de plus grande amplitude que les B-facteurs (Tartaglia et al. 2007). Wiggle enfin est basée sur l'étude des fluctuations à grande échelle observées grâce à une modélisation des protéines par un réseau gaussien. La prédiction utilise les SVMs et l'analyse de données évolutionnaires (Gu et al. 2006).

Une seconde catégorie de méthodes repose sur le concept de *déformabilité*. Ainsi, Boden et collaborateurs se sont appuyés sur DSSPcont (Andersen et al. 2002) (paragraphe 2.2.5), pour développer une méthode capable prédire la probabilité d'appartenir à 8 états différents de structures secondaires (Boden and Bailey 2006). Une mesure d'entropie est dérivée et permet de prédire la classe de flexibilité du résidu, *i.e., variable* (ne change pas de structure secondaire lors de modifications conformationnelles), *non-variable* (change d'assignation). La méthode est testée sur les paires de structures cristallographiques stockées dans la base MolMovDB (paragraphe 5.3.4.2). De même, pour le développement de FlexRP, Chen et collaborateurs ont étudié 66 segments de séquences pour lesquelles plusieurs conformations sont observées dans diverses structures cristallographiques enregistrées dans la PDB. Ils proposèrent ensuite de prédire les fragments présentant une déformation plus importante qu'un seuil donné (Chen et al. 2007a).

Par ailleurs, il faut noter que de nombreuses méthodes sont dédiées à la prédiction du désordre. Les principales sont comparées dans (Ferron et al. 2006) et (Tompa 2008). Leurs prédictions sont parfois utilisées pour évaluer la flexibilité de régions ordonnées (Jin and Dunbrack 2005). Toutefois, leur description détaillée dépasse le cadre de ce manuscrit

.

6. Analyse et Prédiction de la Flexibilité Protéique Locale (manuscrit en cours d'écriture)

6.1 Objectif

Dans cette étude, nous avons étendu notre analyse de la prédiction des structures locales, pour nous intéresser à la notion de *plasticité structurale* d'une séquence et, donc à sa *prédictibilité* d'un point de vue structural. Plusieurs questions ont rapidement émergé : les erreurs de prédiction structurales sont-elles dues à un manque de spécificité de séquence, à une forte flexibilité du fragment considéré ou à une combinaison de ces deux explications ? Les fragments les plus difficiles à prédire sont-ils les plus flexibles ? Et réciproquement, la qualité de la prédiction structurale est-elle un indicateur de la flexibilité d'un fragment ?

Ainsi, nous avons étudié la flexibilité des protéines selon deux points de vue différents : le B-facteur issu des expériences de cristallographie et les fluctuations calculées lors de simulations de dynamique moléculaire (voir paragraphes 5.3.3.1 et 5.3.4.1). Du fait des incertitudes expérimentales pouvant exister en cristallographie et des limites des simulations de dynamique moléculaire, ces deux mesures apportent des informations complémentaires.

Nos résultats nous ont conduits au développement d'une méthode de prédiction de la flexibilité tout à fait originale et reposant sur la prédiction des structures locales. Quelques méthodes de prédiction de la flexibilité s'appuient déjà sur la prédiction des structures secondaires (cf. Tableau 12). Cependant, aucune ne bénéficient des nuances et des spécificités de séquence et d'interaction locales apportées par les alphabets structuraux. Aucune non plus ne tient compte de la qualité des prédictions structurales. Or, la flexibilité est si intimement liée à la séquence et la structure que des informations concernant la prédiction des structures locales pourraient s'avérer intéressantes pour la prédiction de la flexibilité.

Dans cette section, je présenterai tout d'abord une analyse de la relation entre la flexibilité d'un fragment de séquence et la qualité de nos prédictions structurales. Puis, réciproquement, je présenterai notre étude concernant la pertinence de l'information contenue dans la prédiction des structures locales pour le développement d'une méthode de prédiction de la flexibilité.

6.2 Jeu de structures cristallographiques

Un jeu de 172 structures cristallographiques de haute-résolution (≤ 1,5 Å) est extrait de la PDB grâce au service web PDB-REPREDB (Noguchi et al. 2001). Ces protéines partagent moins de 10% d'identité de séquence et diffèrent géométriquement d'un Cα RMSD supérieur à 10 Å. Les protéines plus petites que 70 résidus n'ont pas été conservées. De même, de façon similaire à (Boden and Bailey 2006), les protéines plus grandes que 200 résidus ainsi que celles constituées de plusieurs domaines ou impliquées dans un complexe ont été écartées pour éviter les mouvements à grande échelle. Par ailleurs, pour travailler avec un nombre raisonnable de protéines tout en conservant un jeu de structures représentatif, seules les protéines appartenant aux quatre classes SCOP principales ont été sélectionnées (Tout-α, Tout-β, α/β et α+β) (Murzin et al. 1995). Ces 115 structures ont ensuite été inspectées manuellement pour écarter celles établissant des interactions étendues ou trop internes avec des ligands. En effet, ces derniers pourraient influencer la dynamique propre de la protéine. Finalement, un jeu de 43 protéines a été conservé, *i.e.,* 5 Tout-α, 10 Tout-β, 6 α/β et 22 α+β. Sa composition en structures secondaires (DSSP) est représentative de celle observée dans les protéines connues, *i.e.,* 35,1 % des résidus sont en hélice α, 27,4 en brin β, 19,7 en coude β et 17,8 % en boucles. A titre de comparaison, dans une banque de données non redondante plus grande de 1421 structures (résolution 1,5 Å, identité de séquence < 30 % et Cα RMSD > 10 Å), la distribution des structures secondaires est respectivement 37,8, 21,4, 20,9 et 19,9 %. Ce jeu de 43 structures cristallographique a été assigné en termes de PSLs. Nous verrons au paragraphe 6.3.2 que trois protéines supplémentaires ont du être écartées. Ainsi, 40 protéines seront finalement analysées, soit 4942 fragments chevauchants de 11-résidus.

6.3 Différentes sources de mesure de la flexibilité

6.3.1 B-facteurs cristallographiques

Les B-facteurs ont été extraits pour chaque Cα à partir des fichiers PDB. Comme classiquement réalisé, pour éviter tout biais du à des variations de paramètres expérimentaux (voir paragraphe 5.3.3.1), les B-facteurs bruts ont été normalisés par la méthode préconisée par (Smith et al. 2003). Ainsi, après éviction des valeurs extrêmes détectées statiquement par l'approche de la médiane (*Median-based approach*), les B-facteurs normalisés sont calculés comme suit :

$$\text{B-facteur}_{\text{Norm}} = (\text{B-facteur}_{\text{Bruts}} - \mu)/\sigma$$

Avec μ et σ, la moyenne et la déviation standard des B-facteurs restants. D'un point de vue local, la mobilité d'un fragment de 11 résidus de long sera caractérisée par le B-facteur$_{Norm}$ associé au Cα de son résidu central.

6.3.2 Fluctuations au cours de simulations de dynamique moléculaire

6.3.2.1 Protocole de Simulation

Des simulations de dynamique moléculaires ont été réalisées pour les 43 protéines sélectionnées grâce au logiciel GROMACS 3.3.1 (Lindahl et al. 2001). Le champ de force GROMOS96 43A1 (van Gunsteren et al. 1996) et le modèle d'eau explicite SPC (*Simple Point Charge*) (Berendsen et al. 1981) ont été utilisés.

Avant le lancement de chaque dynamique, les ligands externes restants sont enlevés pour éviter tout biais. Chaque structure est ensuite immergée dans une boite d'eau périodique neutralisée avec des contre-ions Na^+ ou Cl^-. Chaque système est minimisé grâce à l'algorithme de la « plus grande pente » (*Steepest descent* en anglais) durant 1000 pas.

Pour les étapes suivantes, la pression et la température sont maintenues constantes à 300 K et 1 bar grâce à l'algorithme de Berendsen (Berendsen et al. 1984) avec τ_T=0,1 ps and τ_P=0,5 ps. Un pas d'intégration de 2 *fs* a été choisi et la longueur des liaisons covalentes a été contrainte grâce à l'algorithme LINCS (Hess et al. 1997). Un seuil de 1,4 nm a été utilisé pour les interactions non liées et couplé à un algorithme *Generalized-Reaction-Field* pour les interactions électrostatiques à longues distances (constante diélectrique de 54) (Tironi et al. 1995). Pour chaque système :

(i) une simulation de dynamique moléculaire *sous contrainte* a tout d'abord été effectuée pendant 100 *ps*, la protéine reste ainsi fixée pendant que les ions et les molécules d'eau acquièrent des vecteurs de vitesse appropriés.

(ii) Une simulation est ensuite conduite pendant 5 *ns*. Durant la trajectoire, l'évolution du système est enregistrée toutes les picosecondes.

Chacune des simulations a été suivie manuellement, seule la phase de production a été utilisée pour les analyses qui suivent. De plus, trois protéines ne se stabilisant pas en moins de 5 *ns* et/ou perdant leurs structures secondaires régulières ont du être écartées. Finalement, 40 trajectoires, cumulant près de 150 *ns* de simulations, ont été analysées.

6.3.2.2 Mesure de la flexibilité

La mobilité des Cα tout au long des trajectoires de dynamique moléculaire a été analysée *via* le calcul du Cα RMSF (*Root Mean Square Fluctuations*) grâce au logiciel GROMACS 3.3.1. Le Cα RMSF mesure l'amplitude des fluctuations d'un Cα donné tout au long de la trajectoire par rapport à une position moyenne de référence. Préalablement au calcul, toutes les structures d'une trajectoire donnée sont alignées en fonction des Cα. Puis, le Cα RMSF est calculé comme suit :

$$RMSF^{i}_{Norm} = \sqrt{\frac{1}{T}\sum_{t=0}^{T}(\vec{R}^{i}_{t} - \vec{R}^{i}_{moy})^{2}}$$

avec $\vec{R}^{i}_{t}$, les coordonnées du Cα en position *i* dans la structure enregistrée au moment *t* de la trajectoire et $\vec{R}^{i}_{moy}$, les coordonnées moyennes du Cα en position *i*. *T* est le nombre de structures enregistrées (une par *ps*).

De même que pour les B-facteurs, les valeurs de RMSF bruts sont normalisées par protéine. Egalement, de même, d'un point de vue local, la mobilité de chaque fragment de 11 résidus de long peut être caractérisée par le $RMSF_{Norm}$ associé au Cα du résidu central.

6.4 Relation entre la prédiction structurale et la flexibilité

6.4.1 Evaluation de la prédiction des structures locales sur les protéines simulées par dynamique moléculaire

Pour analyser l'influence de la flexibilité du squelette polypeptidique sur la prédiction structurale, une prédiction des structures locales a été réalisée sur le jeu de 40 protéines sélectionné.

Un Q_{120} de 35,7 % et un taux de prédiction (critère géométrique de 2,5 Å) de 61,3 % ont été obtenus. Ces résultats sont proches de ceux obtenus dans notre précédente étude (voir paragraphe 4.1.3.1) (Bornot et al. 2009).

La Figure 61 présente la relation entre les taux de prédiction et les flexibilités moyennes des 120 classes de structures locales. Quelque soit la mesure de flexibilité considérée (B-facteur$_{Norm}$ ou $RMSF_{Norm}$), les classes les plus flexibles sont également les moins bien prédites et inversement. Ainsi, un coefficient de corrélation de Pearson très significatif est

observé entre les taux de prédictions et les B-facteur$_{Norm}$ moyens par classe structurale, *i.e.,* $r_{Pearson}$ = -0,62. De même, en considérant le RMSF$_{Norm}$, la corrélation est de -0,71.

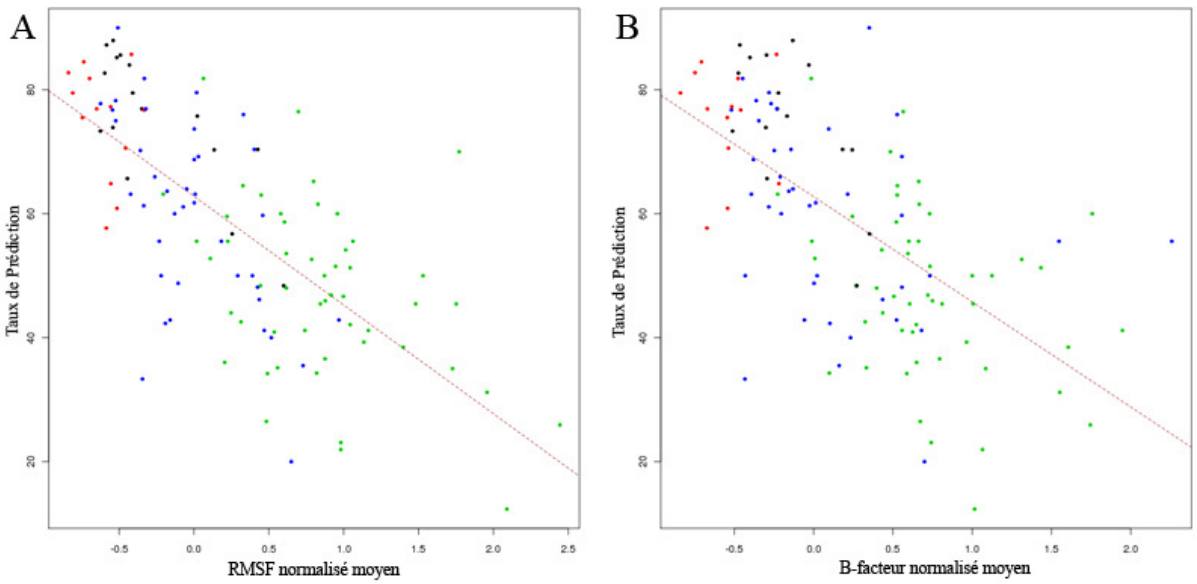

Figure 61. Relation entre la prédiction des structures locales et leur flexibilité.

Chaque point correspond, en abscisse, à la valeur moyenne de flexibilité d'une classe de structure locale donnée et, en ordonnée, au taux de prédiction correspondant. A – La flexibilité est mesurée par le RMSF$_{Norm}$. B – par le B-facteur$_{Norm}$. La couleur des points correspond aux catégories de PSLs proches des structures secondaires, *i.e.,* les PSLs hélicoïdaux, étendus, de connexion et d'extrémité de structures étendues sont en noir, rouge, vert et bleu respectivement. La droite de régression entre les variables est présentée avec une ligne pointillée marron.

Ainsi, les PSLs de connexion restent globalement les plus difficiles à prédire, avec un taux de prédiction de 45,85 % et une précision de 2,57 Å au mieux parmi les 5 candidats. Cette difficulté est généralement couplée à une forte *mobilité* quelque soit la mesure considérée. (voir Figure 61 A et B, les classes structurales vertes). Elle est aussi liée à une forte *déformabilité*. En effet, durant les simulations de dynamique moléculaire, les fragments assignés à une classe de connexion dans les PDB changent en moyenne 3,55 fois de classe structurale (σ = 2,88). Pourtant, les PSLs de connexion sont en moyenne géométriquement différents de plus 3,82 Å des autres PSLs (Cα RMSD). Ainsi, dans la mesure où la diversité *inter*-classe est assez élevée, un changement relativement fréquent d'assignation témoigne donc de deux phénomènes pouvant être concomitants : les fragments des structures de connexion *(i)* subissent de fortes déformations et/ou *(ii)* sont à la frontière de plusieurs classes.

De même, les PSLs caractérisant les extrémités de structures étendues sont le deuxième type de structures les plus difficiles à prédire. Un taux de prédiction toutefois satisfaisant de 61,25 % et une précision de 2,37 Å sont obtenus. Or, durant les simulations de dynamique

moléculaire, les fragments assignés à des PSLs d'extrémités de structures étendues dans les PDB visitent en moyenne 4,58 autres assignations (σ = 3,08). Pourtant, à nouveau, ces PSLs sont en moyenne géométriquement différents de plus de 3,76 Å (Cα RMSD) des autres PSLs. Les catégories de structures locales les mieux prédites sont aussi les moins flexibles. Les structures hélicoïdales et étendues sont associées à 78,7 et 76,5 % de prédictions correctes et des B-facteurs normalisés moyens de -0,14 et -0,55 respectivement (contre de 0,13 et 0,71 en moyenne pour les extrémités de structures étendues et les connexions respectivement).

6.4.2 Relation entre indice de confiance pour la prédiction structurale et flexibilité

La Figure 62 confirme les évolutions inverses de la flexibilité des fragments (B-facteur$_{Norm}$ et RMSF$_{Norm}$) et du taux de prédiction des structures locales en fonction de l'indice de confiance. Les fortes flexibilités (*mobilités*) sont associées à des taux de prédiction faibles, et inversement.

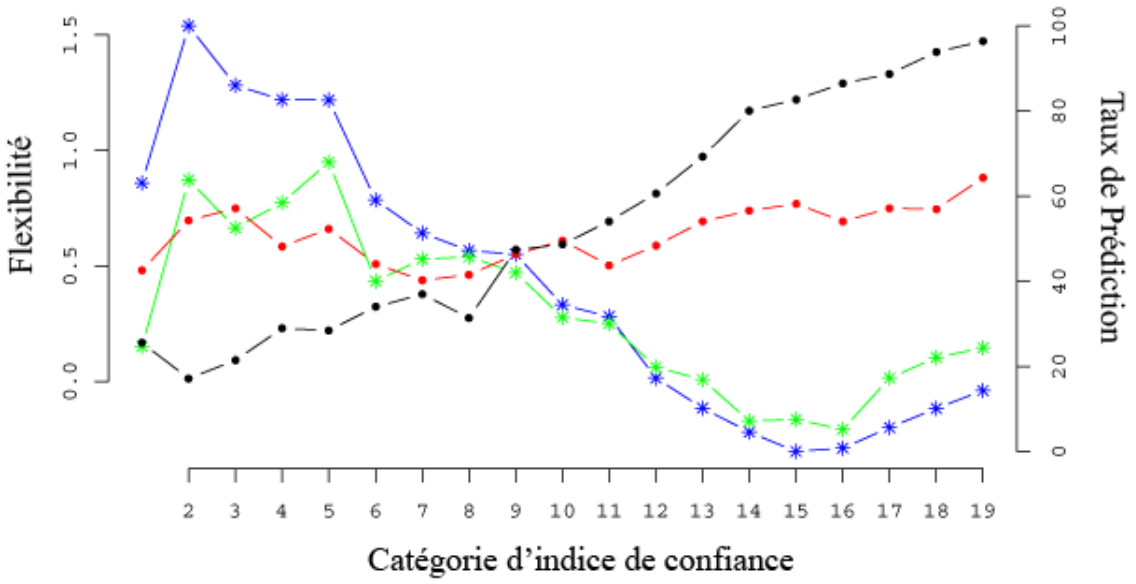

Figure 62. Flexibilité et taux de prédiction en fonction des catégories d'indices de confiance issus de la prédiction des structures locales

L'axe des ordonnées de gauche est dédié aux indices de flexibilité, *i.e.,* RMSF$_{Norm}$ (bleu) and B-facteur$_{Norm}$ (vert) moyens par catégorie d'indices de confiance. L'axe des ordonnées de droite mesure les taux de prédiction obtenus par catégorie d'indice de confiance. Les taux de prédiction des structures locales sont en noir. Les taux de prédiction pour les classes de flexibilité sont en rouge (quadruplet (τ_{B1}, τ_{F1},τ_{B2}, τ_{F2}) = (-1,5 ; -0,5 ; 2,2 ; 1,1), voir paragraphe 6.5.2.4).

Une partie des longs fragments de 11 résidus de long semblent donc ne pas être prédictibles à cause de leur forte flexibilité. Pour la plupart de ces fragments, cette forte *mobilité* s'accompagne d'une forte *déformabilité*. Ainsi, les propriétés dynamiques des protéines devraient être prises en compte dans l'évaluation de la prédiction des structures locales pour dissocier les erreurs réelles des incertitudes dues à une forte flexibilité.

6.4.3 Prise en compte de la dynamique dans l'évaluation des prédictions structurales

Dans un but d'analyse, nous avons pris en compte les simulations de dynamiques moléculaires pour évaluer la prédiction des structures locales. Comme nous l'avons vu paragraphe 3.3.4.1, une prédiction est habituellement considérée comme correcte si l'un des 5 candidats structuraux fournit une approximation meilleure que 2,5 Å par rapport à la structure locale réelle. Cette structure réelle est alors prise comme étant celle observée dans la structure cristallographique. Dans cette analyse, nous considérons à présent que la *vraie* structure réelle est l'ensemble des conformations adoptées par le fragment d'intérêt au cours de la dynamique moléculaire. Ainsi, si nos 5 candidats structuraux fournissent une bonne approximation (≤2,5 Å) pour au moins *x* % de l'ensemble des conformations adoptées durant la simulation, la prédiction est considérée comme correcte. La Figure 63 présente le pourcentage de prédiction habituellement jugées incorrectes devenant correctes en tenant compte des simulations de dynamique moléculaire. En abscisse, x = 50%, signifie que notre prédiction est correcte durant au moins 50 % de la simulation. Dans ces conditions, 9,3 % des prédictions initialement jugées comme incorrectes deviennent correctes, soit 3,6 % de la totalité des fragments prédits.

La Figure 64 présente l'exemple de la prédiction du fragment en position 122-132 dans la structure de l'hydrolase peptidyl-tRNA d'*Escherichia coli* (code PDB 2PTH). Les 5 candidats structuraux prédits sont les PSLs 46, 38, 75, 86 et 31. Le PSLs fournissant la meilleure approximation à 2,63 Å de la structure cristallographique est le PSL n°75. Malgré la forme générale tout à fait satisfaisante de la structure de connexion proposée, cette prédiction est jugée incorrecte (approximation > 2,5Å). Toutefois, pendant la simulation de dynamique moléculaire, le fragment d'intérêt est *mobile* et se *déforme*. Ainsi, nous avons analysé l'évolution de l'approximation donnée par PSL 75 pour cette structure locale flexible tout au

long de la trajectoire de dynamique moléculaire. 33,8 % du temps de simulation, le PSL 75 fournit en fait une approximation satisfaisante.

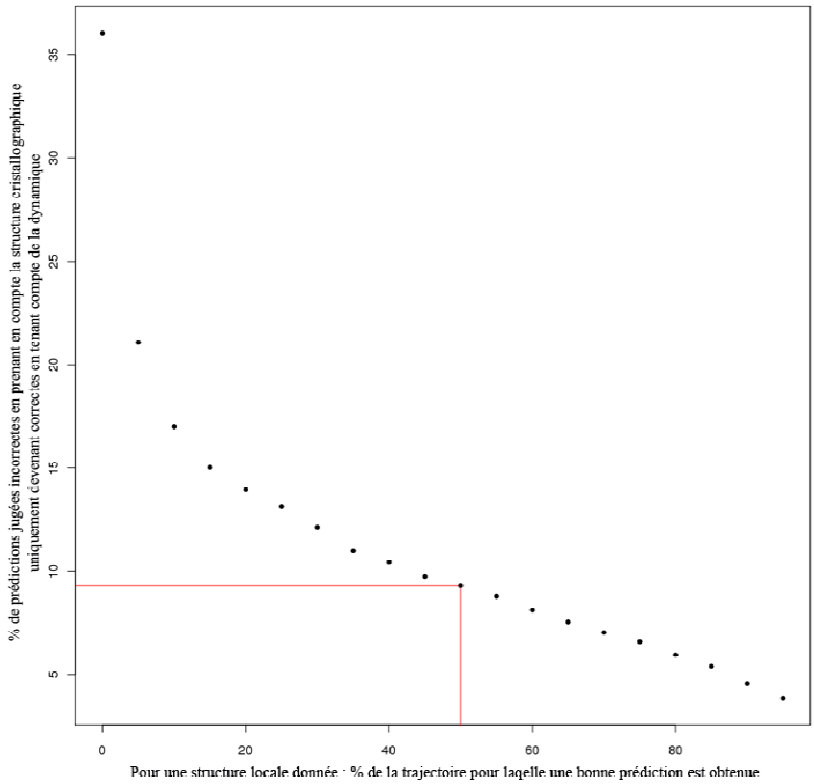

Figure 63. Evaluation de la prédiction des structures locales en tenant compte des simulations de dynamique moléculaire.

% de prédictions jugées incorrectes en prenant en compte uniquement la structure cristallographique et devenant correctes en tenant compte de la dynamique moléculaire, en fonction du % (x) de la trajectoire pour laquelle une bonne prédiction est obtenue. Les lignes rouge indique de si la prédiction doit être bonne 50% de la trajectoire pour être jugée correcte, alors 9,3 % des prédictions initialement jugées mauvaises deviennent satisfaisantes.

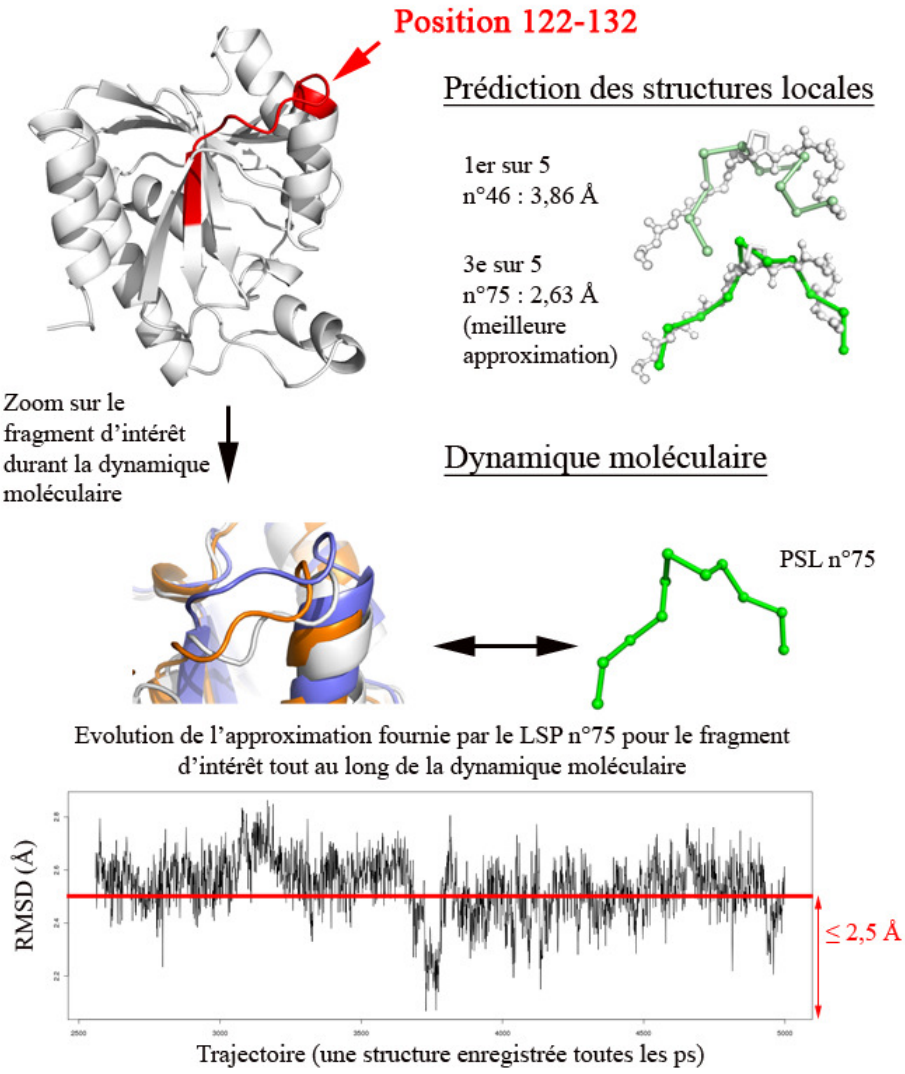

Figure 64. Un exemple de prédiction de structures locales analysé en tenant compte de la flexibilité observée en dynamique moléculaire.

Ces résultats montrent qu'il serait intéressant de prendre en compte la plasticité structurale des fragments de séquence au sein des structures pour évaluer la prédiction des structures locales. Une structure cristallographique ne donne pas accès aux différentes conformations possibles des structures locales flexibles. Or, certains fragments sont trop flexibles pour être correctement représentés par une seule structure. Prédire cette dernière à partir de la séquence n'est donc pas forcément toujours possible puisqu'une même séquence très flexible a pu être figée dans diverses conformations dans différents cristaux.

Réciproquement, les candidats structuraux proposés sont-ils en mesure d'apporter des informations quant à la flexibilité d'un fragment au sein des structures protéiques ? En 2006, Thomas et collaborateurs proposèrent une idée similaire dans le cadre de la prédiction structurale de peptides : une évaluation du polymorphisme structurale est réalisée en se basant

sur la diversité géométrique des ensembles structuraux prédits (Thomas et al. 2006). Dans ce domaine, notre analyse reste ici volontairement assez qualitative car l'étude des *déformations* des structures locales de 11 résidus de long et la comparaison aux différents PSLs prédits pour évaluer leur pertinence, demanderait probablement un échantillonnage plus important que celui que nous avons réalisé. En revanche, la mesure de la *mobilité* locale relative des fragments observée durant nos simulations demande moins de précision et devrait être représentative. Une hypothèse peut donc être vérifiée : les propriétés de *mobilité* des PSLs prédits pour un fragment donné sont elle informatives pour estimer la flexibilité de ce fragment au sein de la structure protéique ?

6.5 Développement d'une méthode de prédiction de la flexibilité

6.5.1 Méthodes

6.5.1.1 Définition de trois classes de flexibilité à partir de deux mesures complémentaires

Les B-facteurs expérimentaux (B-facteur$_{Norm}$) et les fluctuations observées en dynamique moléculaires (RMSF$_{Norm}$) ont été pris en compte simultanément pour définir 3 classes de flexibilité présentées en Figure 65. Une première paire de seuils (τ_{B1}, τ_{F1}) sépare les résidus rigides avec un petit B-facteur et un petit RMSF. Une seconde paire de seuils (τ_{B2}, τ_{F2}) définit la séparation entre les résidus de flexibilité intermédiaire et les résidus avec de forts B-facteur$_{Norm}$ et RMSF$_{Norm}$. . Ainsi, les τ_{Bx} sont les seuils correspondants aux B-facteur$_{Norm}$ et les τ_{Fx} sont les seuils pour les valeurs de RMSF$_{Norm}$. La définition de la valeur du quadruplet (τ_{B1}, $\tau_{F1,}\tau_{B2}$, τ_{F2}) sera présentée et discutée plus loin (paragraphe 6.5.2.1). Les trois classes de flexibilité sont numérotées de 1 à 3, de la plus rigide à la plus flexible.

Chaque Cα du squelette polypeptidique est alors assigné à l'une des trois classes de flexibilité en fonction de ses B-facteur$_{Norm}$ and RMSF$_{Norm}$. De même, chaque fragment de 11 résidus de long, peut être caractérisé par la classe de flexibilité du Cα de son résidu central. Cette classe sera considérée comme leur classe réelle (observée) de flexibilité.

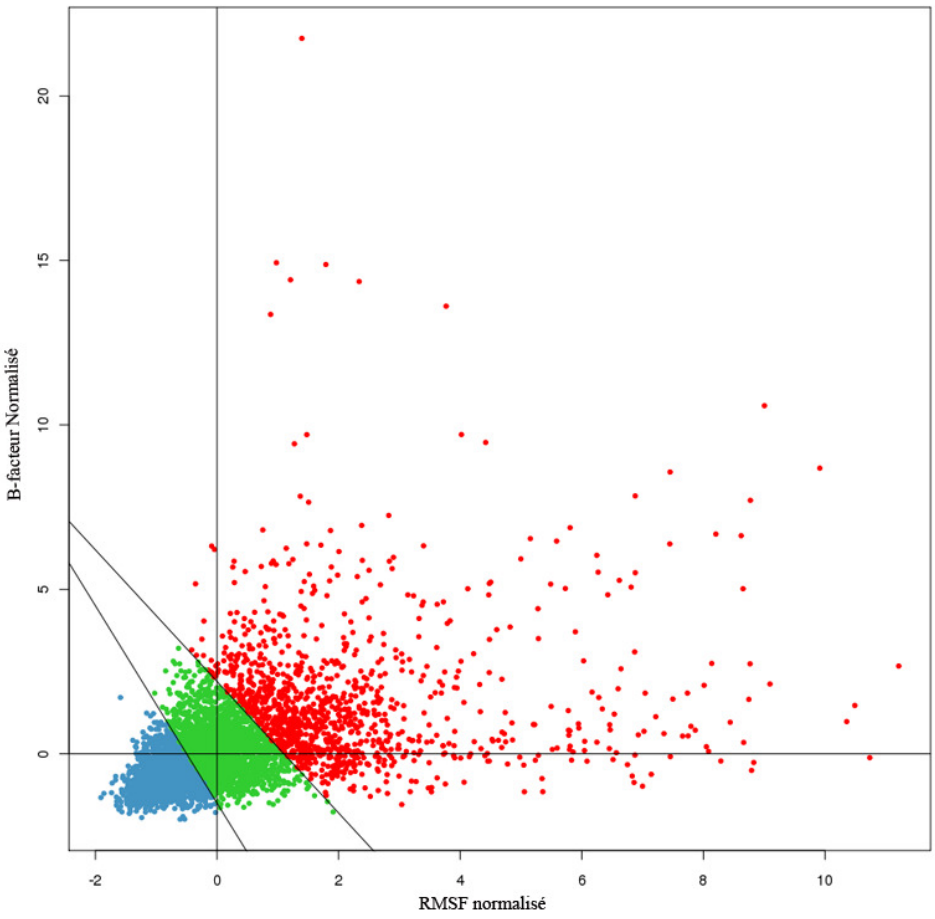

Figure 65. B-facteurs normalisés en fonction des RMSF normalisés issus des simulations de dynamiques moléculaires

Les deux lignes délimitent les trois classes de flexibilité définies par le quadruplet (τ_{B1}, τ_{F1},τ_{B2}, τ_{F2}) = (-1,5 ; -0,5 ; 2,2 ; 1,1) (voir paragraphe 6.5.2.1).

6.5.1.2 Caractérisation des spécificités dynamiques des classes de structures locales

L'objectif ici est d'attribuer une classe de flexibilité préférentielle à chacune des 120 classes structurales. L'attribution ne peut être directe car la plupart des classes contiennent des fragments des trois classes dans des proportions variées. La propension des fragments d'une classe structurale C_s à être associés à une classe de flexibilité C_f a donc été calculée. Elle permet de prendre en compte la distribution des classes de structures au sein des classes de flexibilité :

$$P_{C_f}^{C_s} = \frac{\text{Probabilité}(C_f/C_s)}{\text{Probabilité}(C_f)} = \frac{\frac{n_{C_f}^{C_s}}{n^{C_s}}}{\frac{n_{C_f}}{N}} = \frac{\frac{n_{C_f}^{C_s}}{n_{C_f}}}{\frac{n^{C_s}}{N}}$$ où $n_{C_f}^{C_s}$ est le nombre de fragments assignés à la classe structurale C_s et à la classe de flexibilité C_f. n^{C_s}, n_{C_f} et N sont respectivement le nombre de

fragments assignés à la classe structurale C_s, le nombre de fragments assignés à la classe structurale C_f et le nombre total de fragments. s est numéroté de 1 à 120 et f de 1 à 3. $P_{C_f}^{C_s}$ mesure à quel point les fragments de la classe C_s sont plus associés C_f à la classe qu'attendu par rapport à la banque de données. Finalement, la classe de flexibilité C_f maximisant la propension $P_{C_f}^{C_s}$ est utilisée pour caractériser la classe structurale C_s. De plus, chaque classe structurale sera également caractérisée par la moyenne des B-facteur$_{Norm}$ et RMSF$_{Norm}$ mesurée pour ses fragments, respectivement m_B et m_F.

6.5.1.3 Prédiction de la flexibilité à partir de la prédiction des structures locales

Pour un fragment de séquence donné, notre méthode de prédiction des structures locales propose les 5 candidats structuraux les plus compatibles. Afin d'évaluer l'informativité de ces candidats quant à la flexibilité de la séquence, la prédiction des structures locales est directement utilisée pour obtenir une estimation de la classe de flexibilité d'un fragment. Aucun apprentissage n'est réalisé. La classe de flexibilité estimée correspond simplement à la moyenne arrondie des classes de flexibilité (C_f) des candidats.

Par ailleurs, en suivant le même principe, nous avons également estimé des valeurs théoriques pour les B-facteur$_{Norm}$ et les RMSF$_{Norm}$. Pour chaque fragment de séquence, un B-facteur$_{Norm}$ (RMSF$_{Norm}$) théorique est calculé comme étant la moyenne des m_B (m_F) des 5 candidats. Pour éviter tout biais, la prédiction de ces valeurs théoriques est évaluée par un *jackknife* : pour chacune des protéines p, les valeurs de m_B et m_F sont calculées sur le jeu de structures excluant p et la prédiction est évaluée sur p.

6.5.1.4 Evaluation de la prédiction de la flexibilité

Deux types d'évaluation ont été utilisés pour évaluer la prédictibilité de la flexibilité à partir des structures locales.

La prédiction des classes de flexibilité a tout d'abord été évaluée grâce au calcul du taux de prédiction $Q_3 = VP/N$ avec VP (*Vrais positifs*) le nombre de fragments correctement prédits et N le nombre de fragments total. Par ailleurs, pour permettre une comparaison avec Schlessinger et collaborateurs (Schlessinger and Rost 2005), nous avons également transformé notre prédiction en 3 classes en une prédiction en 2 classes Rigide/Flexible (voir paragraphe 6.5.2.5). Comme ces auteurs, nous avons calculé la F-mesure permettant de combiner la précision (*Accuracy* en anglais) et la sensibilité (*Coverage*) au sein d'une moyenne harmonique:

$$F = \frac{2.ACC.COV}{ACC + COV}$$

Avec *ACC = VP/(VP+FP) et COV= VP/(VP+FN)*. Les *VP, FN* et *FP* sont respectivement le nombre fragments correctement prédits comme Flexible, le nombre de fragments prédits comme Rigides mais observés comme Flexibles et le nombre de fragments prédit comme Flexibles mais observés Rigides.

La deuxième évaluation de l'informativité des structures locales prédites concerne l'estimation des valeurs B-facteur$_{Norm}$ and RMSF$_{Norm}$ théoriques. Le coefficient de corrélation de Pearson entre les valeurs réelles *R* et les valeurs prédites *P* est alors calculé. De plus, comme proposé durant la compétition CASP6 (Jin and Dunbrack 2005), les valeurs *R* ont été divisées en 23 groupes. La corrélation entre la moyenne des *R* pour chaque groupe et la moyenne des *P* correspondante est ensuite calculée. Cette dernière corrélation est moins précise qu'un coefficient de corrélation calculé en prenant compte de chacun des fragments mais permet également de s'affranchir du bruit et de se concentrer sur les tendances.

Il convient de noter que nous ne nous comparerons pas aux résultats obtenus par Chen *et al.* (Chen et al. 2007b). La méthode publiée prédit la flexibilité en 3 classes en se basant sur le B-facteur uniquement. Les auteurs prétendent avoir obtenu les meilleurs résultats observés à ce jour : une corrélation de 0,57 entre les B-facteurs prédits et observés (Tableau 12). Cependant, nous croyons que cette étude est sujette à caution. Plusieurs points semblent en effet problématiques. Par exemple :

- Les auteurs utilisent en entrée de leur méthode d'apprentissage, des informations évolutionnaires HSSP dont la construction est basée sur la connaissance des structures tridimensionnelles et des structures secondaires assignées (Glaser et al. 2005). A l'inverse de toutes les autres méthodes, y compris de la nôtre, la prédiction n'est donc pas réalisée à partir de la seule information de séquence.
- De même, contrairement à toutes les autres études, la normalisation des B-facteurs n'est pas réalisée suivant les préconisations de (Smith et al. 2003).
- De plus, la méthode d'obtention de B-facteurs théoriques (prédits) n'est jamais explicitée.
- Enfin, dans le cadre de la prédiction par classe, les auteurs calculent une précision et une sensibilité pour 3 classes alors que ces calculs sont en théorie définis pour 2 classes seulement. Si les formules sont toutefois appliquées, comme indiqué dans

l'article, dans le cas de 3 classes, les *FN* deviennent égaux aux *FP*. La sensibilité devrait donc être égale à la précision. Or, les auteurs obtiennent deux valeurs différentes.

6.5.1.5 Définition des seuils entre les classes de flexibilité

Etant donné qu'il n'existe aucune limite naturelle permettant de définir aisément des classes de flexibilité. Les frontières entre classes sont souvent choisies arbitrairement (Schlessinger and Rost 2005) ou encore testées de manière systématique et/ou optimisées pour la prédiction (Yuan et al. 2005; Chen et al. 2007b). Nous avons choisi de suivre la seconde stratégie et d'explorer la prédictibilité des classes de flexibilité en fonction de leur définition : la valeur prise par le quadruplet (τ_{B1}, $\tau_{F1,}\tau_{B2}$, τ_{F2}). Ainsi, une grille a été réalisée pour évaluer la prédiction en fonction de la valeur du quadruplet. Les règles suivantes ont été appliquées : *(i)* les seuils τ_{B1}, $\tau_{B2,}\tau_{F1}$, τ_{F2} peuvent prendre toutes les valeurs de l'intervalle $[-2;-0,5] \cup [0,5;5]$ par pas de 0,1, *(ii)* $\tau_{B1} < \tau_{B2}$ et $\tau_{F1}<\tau_{F2}$, *(iii)* une classe de flexibilité contient au minimum 15% des fragments.
Pour chaque quadruplet, la prédiction des classes de flexibilité est évaluée sur tout le jeu de données. Au total plus de 202 977 quadruplets, ont été testé. Ainsi, pour simplifier l'analyse, une stratégie de sélection automatique des meilleurs quadruplets a été utilisée. Nous avons favorisé une prédiction équilibrée pour toutes les classes et non la maximisation du taux de prédiction global (Q_3). Les critères de sélection sont détaillés en Figure 66.

Plus un quadruplet est associé à une prédiction répondant à ces critères, plus il obtient un score élevé (voir Figure 67). Le meilleur quadruplet est finalement sélectionné pour la suite des analyses.

Ainsi, dans le but de simplifier l'analyse et comme réalisé dans (Yuan et al. 2005), le meilleur quadruplet a été choisi sur l'ensemble du jeu de données. Dans le but de vérifier que nous ne surestimons pas la prédiction de la flexibilité en ayant optimisé les classes sur les protéines de test. Nous avons également évalué cette dernière en sélectionnant des quadruplets sur le principe du *jackknife* : le quadruplet est choisi sur toutes les protéines sauf une, cette dernière étant utilisée pour tester la prédiction. 40 tours de *jackknife* ont donc été réalisés (un par protéine). A chaque tour, un nouveau quadruplet est automatiquement choisi et la prédiction évaluée.

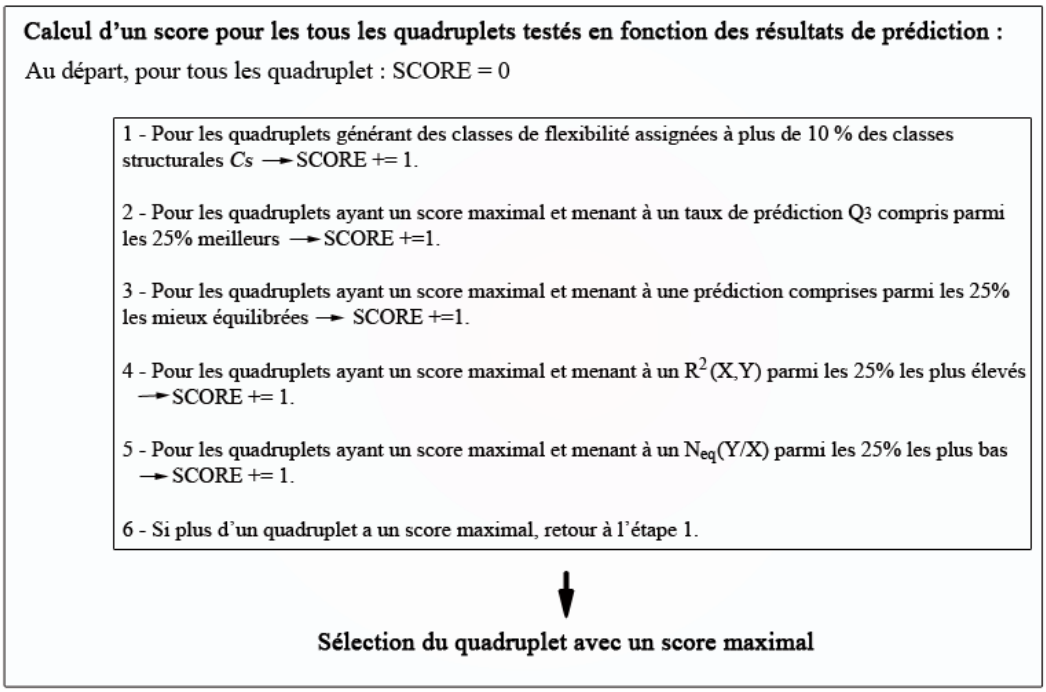

Figure 66. Attribution d'un score à chacun des quadruplets en fonction de la qualité de la prédiction obtenue.

$R^2(X,Y)$ correspond au coefficient de corrélation au carré, il mesure la dépendance non linéaire observé entre les classes observées (X) et prédites (Y). $N_{eq}(Y/X)$ est le nombre équivalent de classes prédites Y sachant la classes observées X (Hazout 2007). Ces deux mesures ont été développées spécifiquement pour l'analyse des prédictions multiclasses.

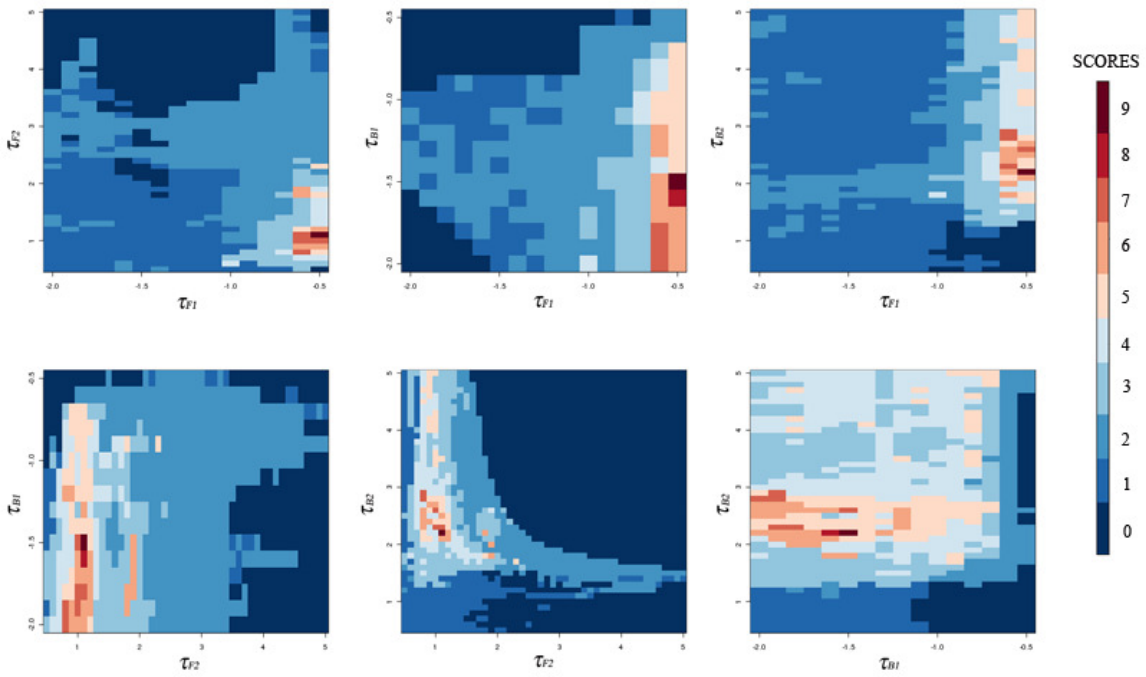

Figure 67. Exploration de l'espace des quadruplets pour la définition des classes de flexibilité les plus prédictibles.

Représentation de la grille d'évaluation réalisée pour tester les différents quadruplets possibles. 6 images sont nécessaires pour présenter les résultats en fonction des seuils deux à deux. Plus un quadruplet obtient un score élevé (prédiction globale de bonne qualité), plus il est rouge.

6.5.2 Résultats

6.5.2.1 Relation entre les deux mesures de flexibilité et analyse des classes

La relation entre les B-facteur$_{Norm}$ expérimentaux et les RMSF$_{Norm}$ issus des simulations de dynamique moléculaire est présentée Figure 65. Le coefficient de corrélation de Pearson entre ces deux mesures est de 0,46 (0,50 en moyenne par protéine, σ = 0,20). Elles présentent donc des tendances similaires. Toutefois, des divergences sont observées. En effet, qualifier un résidu de flexible en fonction du B-facteur peut être considéré comme une erreur en fonction du RMSF et réciproquement. Une vision unifiée prenant en compte plusieurs mesures est nécessaire pour tendre vers les véritables propriétés de flexibilité des structures protéiques.

Ainsi, nous avons pris en compte à la fois le B-facteur$_{Norm}$ et RMSF$_{Norm}$ pour définir trois classes de flexibilité. Comme représenté en Figure 65, une première paire de seuils (τ_{B1}, τ_{F1}) définit la séparation entre les résidus rigides et les intermédiaires. Une seconde paire (τ_{B2}, τ_{F2}) établit la frontière entre les résidus intermédiaires et les résidus flexibles. Le quadruplet de seuils permettant d'obtenir les classes les plus prédictibles est (-1,5 ; -0,5 ; 2,2 ; 1,1) (voir Figure 65 et Figure 67).

Les trois classes de flexibilité définies englobent, de la plus rigide à la plus flexible, 40,4, 36,7 et 22,9 % des fragments de 11 résidus de long. Au sein de la classe rigide, le B-facteur$_{Norm}$ (RMSF$_{Norm}$) est en moyenne de -0,70 (σ = 0,51) (-0,85 (σ = 0,31)). Dans la classe intermédiaire, ces valeurs augmentent : 0,14 (σ = 0,78) et 0,10 (σ = 0,44) respectivement. Enfin, au sein de la classe flexible, le B-facteur$_{Norm}$ et RMSF$_{Norm}$ moyens sont respectivement de 1,62 (σ = 2,05) et 2,12 (σ = 1,80).

Pour quantifier la dispersion des résultats obtenus avec les deux mesures de flexibilité et vérifier l'intérêt de les prendre en compte simultanément, nous avons étudié les résultats que nous aurions obtenus en définissant 3 classes de flexibilité reposant uniquement sur le B-facteur$_{Norm}$ ou sur le RMSF$_{Norm}$. Ces classes sont définies en conservant la proportion des fragments observés dans les classes prenant en compte les deux mesures (cf. ci-dessus). Le Tableau 13 quantifie la confusion entre les 3 classes de B-facteur$_{Norm}$ et celles de RMSF$_{Norm}$. Globalement, la diagonale est assez peuplée et est cohérente avec la corrélation observée entre les deux mesures. Toutefois, la dispersion est également assez importante et, de façon

intéressante est bien équilibrée de chaque côté de la diagonal. Les confusions les plus élevées concernent les fragments rigides et intermédiaires d'une part et, les fragments intermédiaires et flexibles d'autre part. Par exemple, 11,2 % des fragments rigides avec le B-facteur$_{Norm}$ sont intermédiaires avec le RMSF$_{Norm}$ et réciproquement, 11,9 % des fragments intermédiaires avec le B-facteur$_{Norm}$ sont rigides selon le RMSF$_{Norm}$. Il convient toutefois de noter que ces chiffres sont inférieurs à ce qui serait attendu aléatoirement. En revanche, la confusion entre fragments flexibles et intermédiaires est supérieure à l'attendu. Par ailleurs, la confusion entre fragment rigides et flexibles reste assez faible, *i.e.,* 2,8 % et 2,1 % de part et d'autre de la diagonale.

La dispersion des données peut donc être imputée aux deux mesures à part égale. Du côté de la dynamique moléculaire, des imperfections au niveau du paramétrage des champs de force ou encore un défaut d'échantillonnage ont pu mener à des imprécisions. De même, les B-facteurs expérimentaux peuvent également être entachés d'erreurs dues au facteur humain, aux contacts au sein du cristal ou à l'utilisation de petites molécules pour favoriser la cristallisation des régions flexibles en les rigidifiant. Ainsi, il est notamment possible d'obtenir une légère amélioration de la corrélation entre B-facteur$_{Norm}$ et RMSF$_{Norm}$ en ne prenant pas en compte les résidus à moins de 8 Å d'un ligand ($r_{Pearson}$ = 0,49 au lieu de 0,46).

Tableau 13. Confusion entre les B-facteurs normalisés et les RMSF normalisés en trois classes.

		RMSF$_{Norm}$			
		1	2	3	Σ
B-factor$_{Norm}$	1	**26.47**	11.19	2.79	40.45
	2	11.88	**16.03**	8.78	36.69
	3	2.10	9.47	**11.29**	22.87
	Σ	40.45	36.69	22.87	**100.00**

La dernière ligne (colonne) indique la distribution des fragments dans les 3 classes de valeurs de RMSF$_{Norm}$ (B-facteur$_{Norm}$). Le reste du tableau est normalisé de façon à ce que le total fasse 100 %.

Ces observations renforcent l'idée de complémentarité entre les deux visions expérimentale (B-facteur$_{Norm}$) et bioinformatique (RMSF$_{Norm}$). La prise en compte de ces deux mesures devrait permettre de parvenir à une vision plus complète de la flexibilité des structures et à ainsi à une plus grande confiance dans les propriétés mises en évidence. Par exemple, les 22,9% de fragments composant la classe flexible construite selon les mesures sont plus vraisemblablement réellement flexibles dans le contexte de la cellule et pourraient être fonctionnellement importants.

6.5.2.2 Description de la flexibilité des structures locales

Dans le but de caractériser la flexibilité des structures locales, nous avons calculé les B-facteur$_{Norm}$ moyens, m_B, et les RMSF$_{Norm}$ moyens, m_F, par classe de PSL. La Figure 68 présente leur relation. Un coefficient de corrélation de Pearson élevé de 0,77 est observé. Les PSLs les plus rigides selon les deux mesures sont les n° 9 et 10. Ils caractérisent tous deux des structures étendues. Une plus grande dispersion est observée au niveau des classes les plus flexibles : selon le B-facteur$_{Norm}$, les PSLs 89, 87 et 55 sont les plus flexibles tandis qu'avec le RMSF$_{Norm}$, ce sont les PSLs 85, 113 et 103. Toutefois, globalement, même si les ordres de classement sont légèrement différents, les PSLs les plus flexibles le sont pour les deux mesures. Comme attendu, ce sont majoritairement des PSLs de connexion.

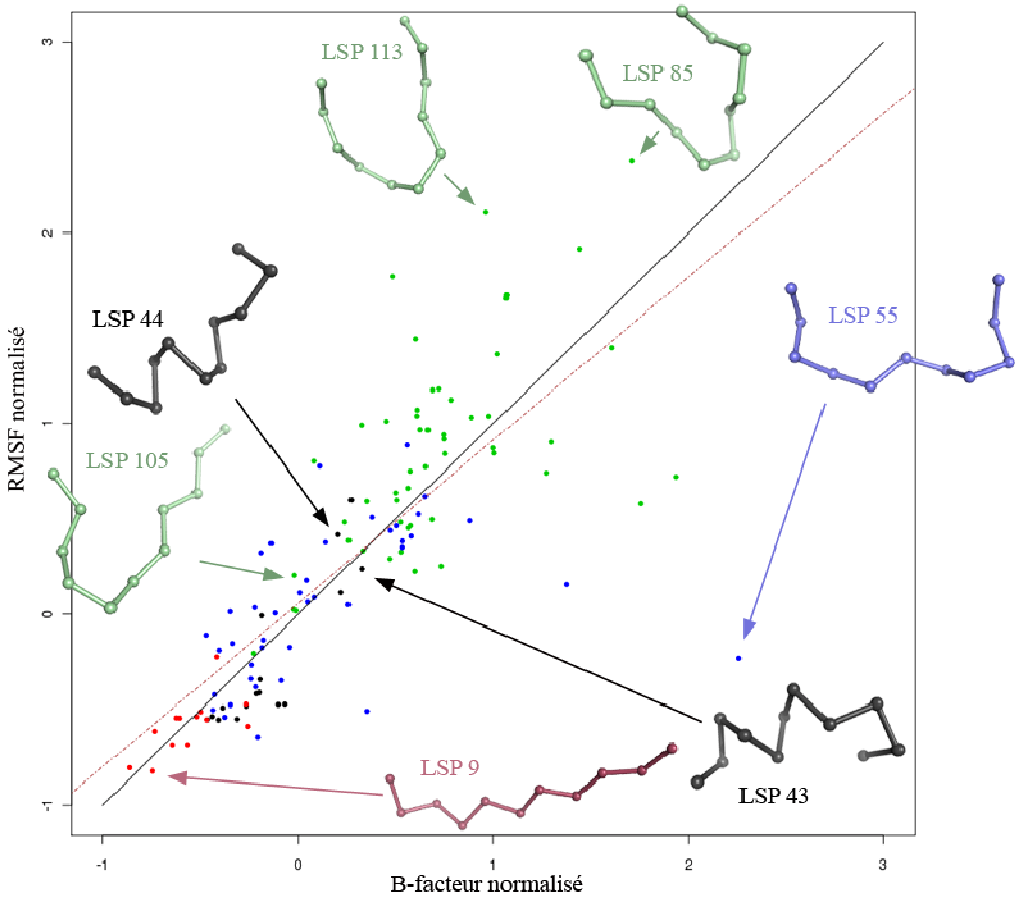

Figure 68. Relation entre les B-facteurs et les RMSF normalisés moyens par classe de PSL.

La couleur des points correspond aux catégories de PSLs proches des structures secondaires, *i.e.* les PSLs hélicoïdaux, étendus, de connexion et d'extrémité de structures étendues sont en noir, rouge, vert et bleu respectivement. La droite de régression entre les variables est présentée avec une ligne pointillée marron. La ligne noire est la première bissectrice.

La Figure 68 semble suggérer que chaque classe de structures locales pourrait être associée à un certain degré de flexibilité, certaines structures de connexion étant par exemple plus *mobiles* que d'autres. Les propriétés spécifiques de chaque classe structurale telles que les interactions locales stabilisant la structure ou les préférences de transition d'un PSL à l'autre au sein des structures (voir Annexe 2) seraient donc prépondérantes. Toutefois, au sein de chaque classe structurale, de fortes déviations standards par rapport aux moyennes m_B et m_F sont observées, *i.e.* σ = 1,29 et 1,23 en moyenne pour les B-facteur$_{Norm}$ et les RMSF$_{Norm}$ respectivement. Cette dernière observation montre que d'autres facteurs comme les interactions à longues distances, sont également impliqués dans la flexibilité des fragments. Est-il tout de même possible d'affirmer que les classes de structures locales ont des propriétés de flexibilité spécifiques ?

Nous avons étudié les distributions des valeurs de flexibilité au sein de chaque classe structurale. Un test de comparaison statistique a ainsi été effectué entre les différentes classes prises deux à deux. En raison du faible peuplement de certaines classes structurales[6], le classique *test-t* paramétrique n'a pu être réalisé, les conditions de validité n'étant pas réunies. Un test non-paramétrique de Mann-Whitney-Wilcoxon a donc été choisi. La Figure 69 présente les *p-values* obtenues pour toutes les comparaisons effectuées sur les valeurs de B-facteur$_{Norm}$. Ce test permet de voir si deux classes structurales sont associées à des distributions de B-facteur$_{Norm}$ (RMSF$_{Norm}$) proches ou non.
Cette analyse montre que les B-facteurs d'une classe structurale donnée sont en moyenne significativement différents de ceux de 58,9 autres classes (*p-value* < 0,05). La même analyse a été réalisée en prenant en compte le RMSF$_{Norm}$. Une classe structurale donnée devient alors significativement différentes de 68,1 autres classes en moyenne. Ces observations supportent l'existence d'une certaine spécificité des propriétés de flexibilité des différentes classes de structures locales. Ainsi, la *mobilité* de la classe de connexion 85 est par exemple trouvée significativement différente de celles des classes 105 ou 30, également considérées comme des boucles (cf. Figure 70).

En conséquence, même si elles n'expliquent pas en totalité la flexibilité du squelette polypeptidique, les structures locales adoptées par ce dernier semblent avoir une forte influence.

[6] Rappelons que les analyses sont réalisées sur 40 protéines.

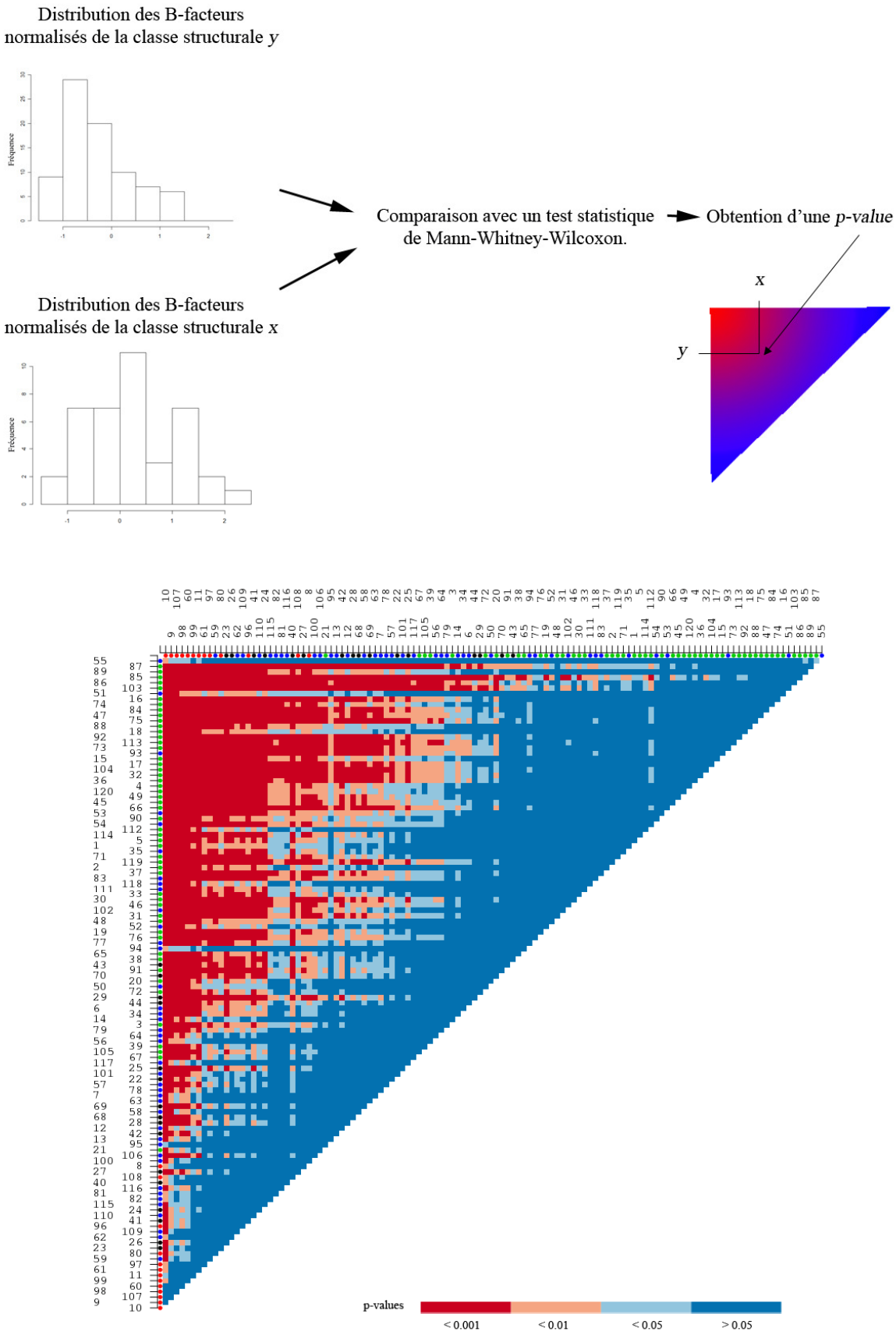

Figure 69. P-values obtenues à partir des tests statistiques de Mann-Whitney-Wilcoxon comparant les distributions de B-facteurs normalisés observées pour les 120 classes structurales.

Légende de la figure 69

La partie haute de la figure explique le principe de l'obtention de la demi-matrice en bas.
Haut : Les distributions des B-facteur$_{Norm}$ dans les classes structurales *x* et *y* sont comparées. Un test statistique de Mann-Whitney-Wilcoxon est utilisé et une *p-value* est calculée (logiciel R, (Ihaka and Gentleman 1996)). Cette dernière mesure la probabilité d'obtenir une statistique aussi extrême que celle observée sous l'hypothèse nulle (= les deux distributions sont similaires). Ainsi, plus la *p-value* est faible, plus les distributions sont significativement décalées.
Bas : La demi-matrice présente les *p-values* obtenues pour toutes les comparaisons entre les 120 classes prises deux à deux. Elles sont colorées du rouge au bleu, des plus significatives au moins significatives. Les classes structurales en abscisse *(x)* et en ordonnée *(y)* sont ordonnées de la même manière, de la plus rigide à la plus flexible selon les moyennes m_B de B-facteur$_{Norm}$. A côté du numéro de chaque classe, un point indique la catégorie de structure secondaire de chaque classe (voir Figure 68).

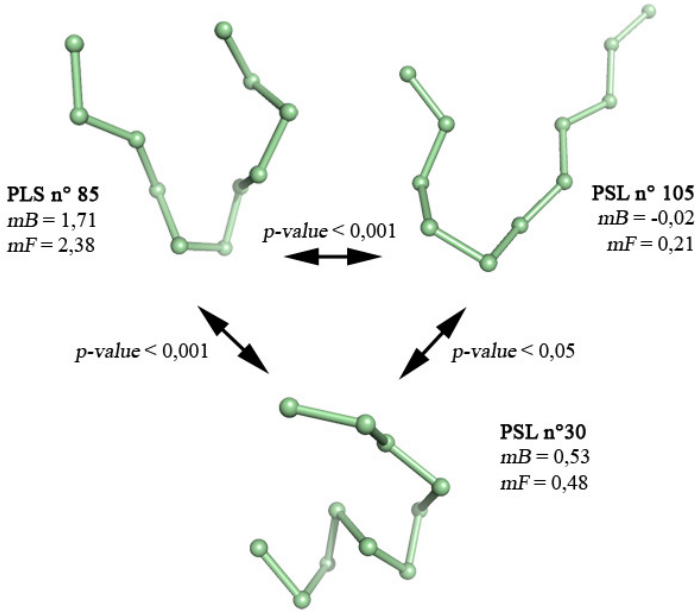

Figure 70. Exemples de PSLs de connexion dont les degrés de flexibilité sont significativement différents.
Les PSL n°85 et 30 appartiennent à la classe flexible, le PSL n°105 appartient à la classe intermédiaire.

6.5.2.3 Classification des structures locales en fonction de leur flexibilité

Comme nous l'avons vu dans le paragraphe 6.5.1.2, les 120 classes de structures locales ont été réparties dans les 3 classes de flexibilité en fonction de la propension de leurs fragments à appartenir à la classe rigide, intermédiaire ou flexible. Rappelons que ces classes de flexibilité ont été déterminées en tenant compte à la fois du B-facteur expérimental et des fluctuations observées en dynamique moléculaire.

Ainsi, 35,8 % des PSLs ont été assignés à la classe rigide, 25,0 % à la classe intermédiaire et 39,2 % à la classe flexible. La totalité des PSLs étendus et 62,5% des PSLs hélicoïdaux sont assignés à la classe rigide. Toutefois, 25,0% des classes hélicoïdales sont également assignées à la classe intermédiaire et 12,5% (soit 2 PSLs/16) à la classe flexible. Les deux PSLs hélicoïdaux les plus *mobiles* sont les n° 43 et 44, caractérisant des extrémités C-terminales d'hélices. Ce résultat est en accord avec des études observant moins de contraintes ou une plus grande flexibilité au niveau des extrémités C-terminales des hélices par rapport aux extrémités N-terminales (Miick et al. 1993; Ho et al. 2003). Par ailleurs, aucun PSL de connexion n'est assigné à la classe rigide. 25% sont intermédiaires et 74,5% sont flexibles. Enfin, les PSLs d'extrémité de structure étendues sont à 50,0 % classés comme rigides et à 32,5% comme intermédiaires.

6.5.2.4 Prédiction de la flexibilité

Notre objectif est d'évaluer l'informativité de la prédiction des structures locales pour l'estimation de la flexibilité du squelette polypeptidique. Nous avons mis en place une stratégie pour déduire directement la classe de flexibilité à partir de celles des candidats structuraux prédits, *i.e.,* la flexibilité prédite est la moyenne des flexibilités des candidats (paragraphe 6.5.1.3). Aucun apprentissage spécifique n'a été réalisé.

Afin de prendre en compte le plus de données possible et d'augmenter la représentativité des classes de flexibilité, leurs frontières ont été choisies en fonction du jeu de structures entier. Ainsi les classes les plus prédictibles sont définies par le quadruplet (τ_{B1}, $\tau_{F1,}\tau_{B2}$, τ_{F2}) = (-1,5 ; -0,5 ; 2,2 ; 1,1). Pour nous assurer de la robustesse des résultats dans le cas d'une nouvelle protéine cible totalement inconnue. Nous avons également effectué une sélection de quadruplets selon le principe du *jackknife*, *i.e.,* choix du quadruplet sur toutes les protéines sauf une, évaluation de la prédiction sur cette dernière. Le quadruplet obtenu est en moyenne (τ_{B1}, $\tau_{F1,}\tau_{B2}$, τ_{F2}) = (-1,4 $_{(\sigma=0,3)}$; -0,7 $_{(\sigma=0,2)}$; 2,3 $_{(\sigma=0,4)}$; 1,4 $_{(\sigma=0,5)}$). Les résultats de prédiction, très similaires à ceux obtenu avec le quadruplet défini globalement, seront donnés en parallèle dans les différents tableaux.

6.5.2.4.1 Prédiction de la flexibilité en 3 classes

Un taux de prédiction de 50,9 % est obtenu. La prédiction est très bien équilibrée : les classes rigide (1), intermédiaire (2) et flexible (3) sont prédites avec des taux de 50,4, 51,6 et 50,7 % de prédictions correctes (cf. Tableau 14).

Tableau 14. Matrice de confusion entre les classes de flexibilité assignées et prédites.

		Global parameterization				*Jackknife*			
		Predicted flexibility classes				Predicted flexibility classes			
		1	2	3	Σ	1	2	3	Σ
Assigned	1	**50.44**	38.1	11.46	100.00	**47.43**	39.08	13.49	100.00
Flexibility	2	20.07	**51.61**	28.32	100.00	20.16	**48.34**	31.50	100.00
classes	3	6.17	43.11	**50.72**	100.00	5.80	39.17	**55.03**	100.00
Σ		29.14	44.22	26.63	100.00	26.68	43.14	30.18	100.00

La dernière ligne présente les proportions de fragments prédits dans chaque classe de flexibilité.

Une prédiction des PSLs aléatoire, comme présentée paragraphe 3.3.4.1, aurait mené à un taux de prédiction de seulement 36,0 %. De plus, du fait de la stratégie employée, *i.e.,* une simple moyenne, 79,0 % des fragments seraient prédits en classe intermédiaire. Ainsi, seulement 8,5 % des fragments rigides et 13,8 % des fragments flexibles seraient bien prédits. Nos résultats sont donc largement supérieurs à une prédiction aléatoire.

Par ailleurs, il convient de noter que la confusion vient majoritairement de la classe intermédiaire. En effet, une faible confusion existe entre les classes flexibles et rigides : seulement 11,5 % des fragments rigides sont prédits comme flexibles et 6,2 % des fragments flexibles sont prédits rigides (voir Tableau 14). Cette confusion n'est pas significativement différente de celle observée en comparant des classes de B-facteurs expérimentaux à des classes de RMSF issus de la dynamique moléculaire : 6,9 % des fragments rigides selon le B-facteur$_{Norm}$ sont vus comme flexibles selon le RMSF$_{Norm}$ et réciproquement 9,2 % des fragments flexibles avec le B-facteur$_{Norm}$ sont vus comme rigides selon le RMSF$_{Norm}$.

Enfin, l'évolution du taux de prédiction de la flexibilité en fonction de l'indice de confiance de la prédiction des structures locales est présentée Figure 62. Quelque soit la valeur de l'indice de confiance et le taux de prédiction des structures locales, la prédiction de la flexibilité reste globalement très stable. Toutefois, des variations par classe sont observées. La classe rigide est mieux prédite pour les indices de confiances élevés : un fort indice de confiance (catégories 13 à 19) correspond à un taux de prédiction de 63,02 %, tandis qu'un faible indice de confiance (catégories 1 à 6) correspond à un taux de prédiction de 10,4 %. A l'inverse, la classe flexible est mieux prédite pour les indices de confiance faibles : 63,9 % de prédiction correctes contre 43,7 % au niveau des indices élevés. Ces variations sont en partie liées à la répartition des fragments, seulement 5,3 % des fragments rigides sont associés à des indices de confiance faibles et, de même, bien qu'à un degré moins important, seulement 28,4 % des fragments flexibles sont associés à un fort indice. Toutefois, le taux de prédiction élevé

pour la classe flexible au niveau des indices de confiances faibles montre également que le manque d'informativité de séquence pour la prédiction des structures locales est en fait informatif pour la prédiction de la flexibilité.

6.5.2.4.2 Prédiction de profils de flexibilité

Pour aller plus loin, des valeurs de B-facteur$_{Norm}$ et RMSF$_{Norm}$ ont également été estimées à partir de la prédiction des structures locales. Rappelons que, pour un fragment de séquence cible, notre estimation correspond à la moyenne des flexibilités associées aux 5 candidats structuraux (*i.e.* les m_B pour le B-facteur$_{Norm}$ et les m_F pour le RMSF$_{Norm}$) (voir paragraphe 6.5.1.3). Ainsi, sans aucun apprentissage, nous obtenons une corrélation de Pearson de 0,30 entre les B-facteur$_{Norm}$ observés et les B-facteur$_{Norm}$ prédits. De même, pour les RMSF$_{Norm}$, une corrélation de 0,41 est obtenue. En ne considérant pas les mesures extrêmes pouvant être dues à des biais expérimentaux ou de simulation, ces corrélations valeurs de flexibilité observées et prédites augmentent légèrement et atteignent 0,33 et 0,45 respectivement.

6.5.2.5 Comparaison avec d'autres méthodes

6.5.2.5.1 Prédiction de classes de flexibilité

Nous avons tout d'abord comparé nos résultats de prédiction par classe à ceux obtenus par Schlessinger et al avec leur méthode PROFbval (Schlessinger and Rost 2005). La comparaison ne peut être directe puisque leur prédiction est réalisée pour deux classes de flexibilité (Rigide/Flexible) reposant uniquement sur le B-facteur normalisé (Tableau 12). Deux types de classes ont été définis selon des seuils dits *strict* et *non-strict*, *i.e.* 0,03 et -0,3 respectivement.

Dans un premier temps, nous avons ajusté notre étude en fonction de leur définition des classes. Ainsi, la flexibilité a été décrite en deux classes en fonction du B-facteur$_{Norm}$ uniquement et selon leurs seuils. L'évaluation de la prédiction a alors été réalisée non seulement sur notre banque de 40 protéines mais également sur jeu plus grand de 1041 structures cristallographiques non redondantes (résolution meilleure que 2 Å, identité de séquence inférieure à 30 % et une distance géométrique de plus de 10 Å). Le premier jeu sera nommé le *jeu de MD* (pour *Molecular Dynamics*), le second sera nommé le *jeu de Validation*. Le Tableau 15 montre que nous obtenons une F-mesure de 48,1 et 72,0 % respectivement pour une définition *stricte* et *non-stricte* des classes sur le *jeu de MD*. Ces valeurs sont de 44,9

et 69,7 % sur le grand *jeu de Validation*. PROFbval atteint des F-mesures de 53,3 et 71,9 %. Ainsi, sans apprentissage, nos résultats sont moins précis que ceux de PROFbval selon la définition *stricte* des classes mais tout à fait comparables dans le cas de la définition dite *non-stricte*.

Dans un second temps, nous avons transformé notre prédiction en trois classes en une prédiction en deux classes en utilisant les paramètres de notre étude. La flexibilité est définie à la fois par le B-facteur$_{Norm}$ et le RMSF$_{Norm}$. Le seuil dit *strict* sera la limite entre notre classe rigide et notre classe intermédiaire, *i.e.,* les classes intermédiaires et flexibles sont fusionnées. Le seuil *non-strict* sera la limite entre la classe intermédiaire et la classe flexible, *i.e.,* nos classes rigides et intermédiaires sont fusionnées. Il convient de noter que dans ces conditions, seule la situation *non-stricte* peut être utilisée pour la comparaison. En effet, la répartition des fragments dans chaque classe dans la situation *stricte* est trop différente de la répartition utilisée par PROFbval. Ainsi, dans la situation *non-stricte*, nous obtenons une précision de 71,8 %, une sensibilité de 85,2 % et une F-mesure de 71,9 %. Ces résultats sont légèrement supérieurs à ceux de PROFbval et donc très prometteurs.

Tableau 15. Notre prédiction de la flexibilité transformée pour être comparée à PROFbval.

		This study, adapted for comparison				PROFbval
		-MD dataset -Global parameterization -Bfact$_{Norm}$ and RMSF$_{Norm}$	*-MD dataset -Jacknife -Bfact$_{Norm}$ and RMSF$_{Norm}$*	*-MD dataset - PROFbval thresholds - Bfact$_{Norm}$ Only*	*-Wide validation set - PROFbval thresholds - BfactNorm Only*	
Strict	*threshold value(s)*[a]	2.2 ; 1.1	2.3±0.4 ; 1.4±0.5	0.03	0.03	0.03
	% of rigid fragments	77.17	78.76	57.80	58.20	57.80[b]
	% of flexibile fragments	22.82	21.24	42.20	41.80	42.20[b]
	ACC	43.47	38.73	59.30	55.30	63.1
	COV	50.72	55.03	40.43	37.81	46.2
	F-measure	46.82	45.46	48.08	44.91	53.3
Non-Strict	*threshold value(s)*[a]	-1.5 ; -0.5	-1.4±0.3 ; -0.7±0.2	-0.3	-0.3	-0.3
	% of rigid fragments	40.33	35.09	44.33	45.64	44.33[b]
	% of flexibile fragments	59.67	64.91	55.67	54.36	55.67[b]
	ACC	71.79	74.84	61.23	58.49	70.1
	COV	85.25	84.54	87.35	86.16	73.9
	F-measure	77.94	79.39	71.99	69.68	71.9

[a]: for this study, a threshold line is defined by a B-factorNorm and a RMSFNorm value respectively

[b]: according to our MD dataset

6.5.2.5.2 Prédiction de profil*s de flexibilité*

Comme nous l'avons vu au paragraphe 6.5.2.4.2, nous obtenons une corrélation de 0,30 (0,41) entre les B-facteur$_{Norm}$ (RMSF$_{Norm}$) observés et prédits. La corrélation est de 0,33 (0,45) si les valeurs extrêmes ne sont pas prises en compte. Schlessinger *et al.*, Yuan *et al.* et Zhang *et al.* obtiennent respectivement 0,44, 0,53 et 0,55 pour le B-facteurs sur leur jeu de données (Schlessinger and Rost 2005; Yuan et al. 2005 ; Zhang et al. 2009) (Tableau 12). A nouveau notre objectif n'est pas d'entrer en compétition avec ces méthodes reposant sur des méthodes d'apprentissage sophistiquées. La prédiction de valeurs théoriques est plus complexe que la prédiction de 3 classes. Néanmoins, nous avons étudié l'utilisation de notre stratégie pour la prédiction de valeurs moyennes de B-facteurs regroupées en 23 catégories. Cette analyse a été proposée dans le cadre de la compétition CASP6 pour évaluer la capacité des méthodes de prédiction du désordre à prédire également les B-facteurs normalisés (Jin and Dunbrack 2005) (voir paragraphe 6.5.1.4). La meilleure corrélation avait été obtenue par PONDR VSL1 et atteignait 0,92 (Obradovic et al. 2005). Dans les mêmes conditions, nous obtenons une corrélation de 0,71 pour le B-facteur$_{Norm}$ et de 0,69 pour RMSF$_{Norm}$. Sans prendre en compte les valeurs extrêmes, ces corrélation atteignent même 0,94 et 0,96. Les structures locales prédites pour un fragment de séquence donné sont donc suffisamment informatives pour nous permettre d'identifier directement et de façon satisfaisante les régions plus ou moins flexibles selon une échelle de 23 degrés.

6.5.2.6 Un exemple de prédiction

La Figure 71 présente les profils de flexibilité observés et prédits d'une protéine de liaison aux acides gras de l'intestin de rat (PDB code 1IFC (Scapin et al. 1992)). La corrélation entre les valeurs de B-facteur$_{Norm}$ (RMSF$_{Norm}$) observées et prédites est de 0,43 (0,60). Sans prendre en compte les valeurs extrêmes, elle atteint 0,53 (0,67).

La Figure 72 permet de visualiser sur sa structure la flexibilité observée et prédite de la protéine de liaison aux acides gras.

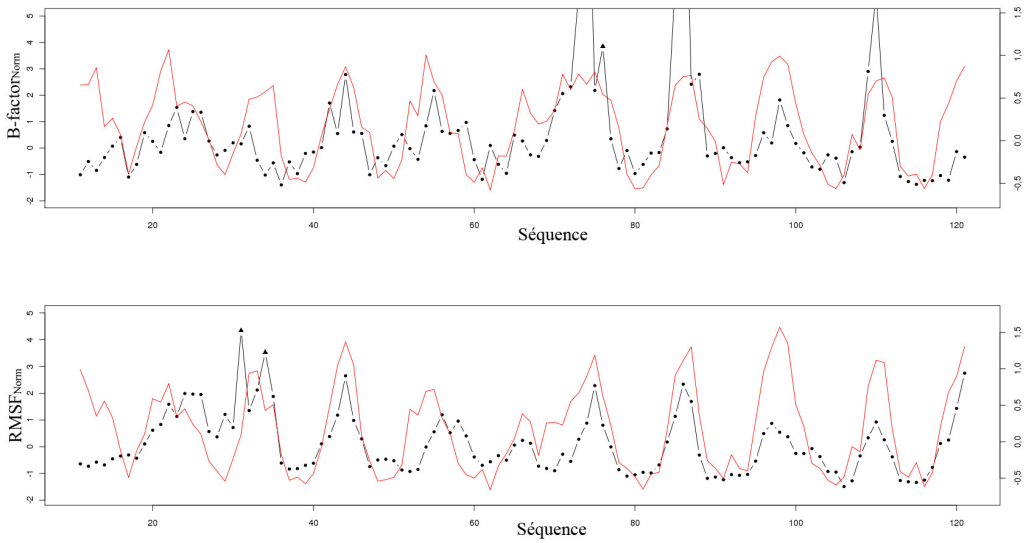

Figure 71. Prédiction de la flexibilité d'une protéine de liaison aux acides gras de rat (code ODB 1IFC, 131 résidus).

Haut : Superposition des valeurs de B-facteur$_{Norm}$ observés et prédits. *Bas* : Superposition des valeurs de RMSF$_{Norm}$ observés et prédits. Les observations sont en noir et les prédictions en rouge. Les valeurs extrêmes (selon l'approche basée sur la médiane utilisée par (Smith et al. 2003)) sont symbolisées par des triangles.

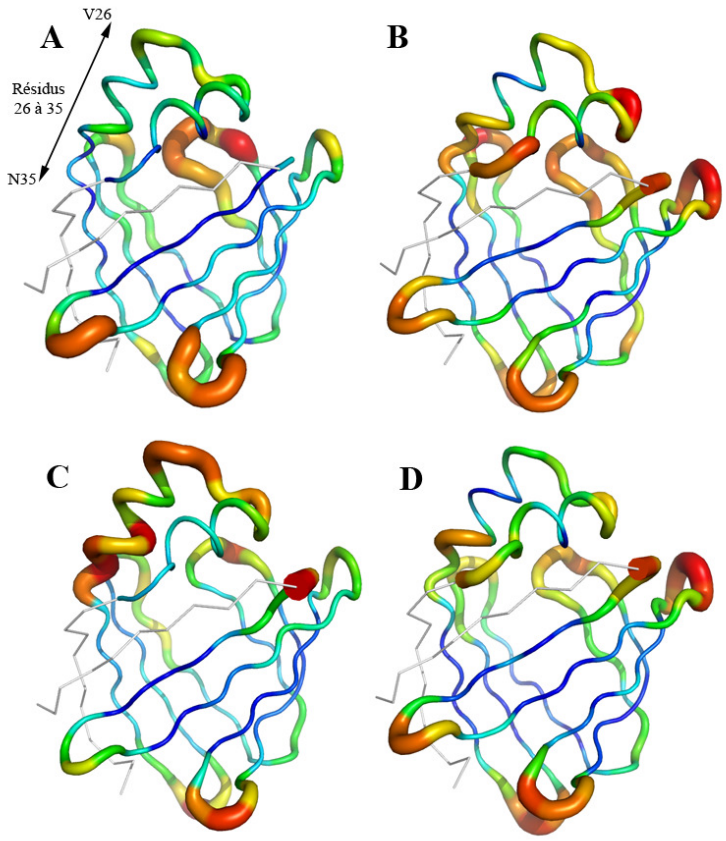

Figure 72. Flexibilité observée et prédite présentées sur la structure d'une protéine de liaison aux acides gras de rat (code PDB 1IFC).

A. B-facteur$_{Norm}$ observés, B. B-facteur$_{Norm}$ prédits, C. RMSF$_{Norm}$ observés, D. RMSF$_{Norm}$ prédits. La position des résidus 26 à 35 est indiquée, voir le texte pour explication.

L'accord entre les valeurs observées et les valeurs prédites est donc assez satisfaisant. En revanche, l'accord entre les deux valeurs observées, B-facteur$_{Norm}$ et RMSF$_{Norm}$, est plus faible, *i.e.* une corrélation de 0,37 est observée. Les résidus 33 à 35, par exemple, sont observés comme flexibles selon le RMSF$_{Norm}$ mais pas selon le B-facteur$_{Norm}$. Toutefois, les valeurs de B-facteur$_{Norm}$ et RMSF$_{Norm}$ prédites sont assez élevées et indiquent donc une zone flexible. De façon intéressante, ces résidus font partie d'une région proposée comme étant un "portail" pour les ligands vers la cavité enfouie dans le cœur de la protéine (Scapin et al. 1992; Friedman et al. 2006). Hodsdon and Cistola (Hodsdon and Cistola 1997) ont montré par RMN que la région allant des résidus V26 à N35 était caractérisée par de faible paramètres d'ordre dans la forme non-liée de la protéine mais pas dans sa forme liée[7]. Ainsi, même si le mécanisme d'entrée des acides gras n'est pas encore connu avec certitude (Friedman et al. 2006), ces résidus pourraient jouer un rôle important dans la fonction de la protéine. Notre prédiction semble donc ici pertinente.

6.6 Extension de notre analyse de la flexibilité et amélioration de la prédiction

Des analyses sont actuellement en cours pour étendre cette étude et améliorer la prédiction de la flexibilité.

6.6.1 Analyse de la flexibilité sur des jeux de données plus grands et prise en compte d'une troisième mesure de flexibilité

Nous avons, tout d'abord, étendu notre analyse pour prendre en compte des protéines plus longues que 200 résidus. Des simulations de dynamique moléculaire ont ainsi été réalisées sur 29 protéines supplémentaires. Notre jeu de dynamique moléculaire est donc à présent composé de 69 protéines, soit 256 *ns* de simulations analysables.

De plus, nous nous sommes appuyés sur la base de données RECOORD pour créer un jeu de modèles RMN. RECOORD contient 545 modèles RMN raffinés en utilisant un protocole standardisé à partir des contraintes originales (Nederveen et al. 2005). Nous avons tout d'abord sélectionné les modèles permettant de conserver moins de 25% d'identité de séquence dans le jeu de données. Puis, dans un second temps, nous avons utilisé l'algorithme FindCore pour écarter les modèles présentant des mouvements de domaines à large échelle et pouvant biaiser notre analyse (Snyder and Montelione 2005). De plus, une attention

[7] La protéine analysée par Hodsdon and Cistola Hodsdon, M.E., and Cistola, D.P. 1997. Ligand binding alters the backbone mobility of intestinal fatty acid-binding protein as monitored by 15N NMR relaxation and 1H exchange. *Biochemistry* **36:** 2278-2290. présente 100% d'identité de séquence avec 1IFC.

particulière a été portée sur la qualité des modèles en considérant les données de contraintes disponibles dans la BioMagResBank (Nederveen et al. 2005). Finalement, un jeu RMN de 368 modèles ont été conservés. Chaque modèle est constitué de 25 structures. Les fluctuations des Cα sur les 25 structures sont alors mesurées en calculant un RMSF comme expliqué pour la dynamique moléculaire (voir paragraphe 6.3.2.2).

Les compositions de ces jeux de données en termes de protéines sont assez différentes. Par exemple, le jeu de RMN contient seulement 4,9 % de protéines de classe SCOP α/β contre 20,3 % dans le jeu utilisé pour la dynamique moléculaire. De même, les protéines de jeu de RMN contiennent plus de boucles, *i.e.* 21,4 % contre 18,8 %. Or, nous avons observé que la forte corrélation observée entre B-facteur$_{Norm}$ et RMSF$_{Norm}$ moyens par PSL sur le jeu de 40 protéines est retrouvée quelque soit le jeu de données. En considérant les protéines plus longues que 200 résidus, la corrélation entre B-facteur$_{Norm}$ et RMSF$_{Norm}$ moyens par PSL pour le jeu de dynamique moléculaire augmente et passe à 0,94. Cette augmentation est majoritairement due à un peuplement plus important des classes. De même, la corrélation entre les RMSF$_{Norm}$ moyens calculés avec les données de simulation et les RMSF$_{Norm}$ moyens issus de RMN est de 0,95. Enfin, la corrélation entre les B-facteur$_{Norm}$ et les RMSF$_{Norm}$ moyens issus de RMN est de 0,89. Ainsi, les propriétés de flexibilité des PSLs semblent conservées quelque soit les protéines et quelques soit le système de mesure.
Des analyses complémentaires montrent de plus que les propriétés d'accessibilité et le nombre de contacts[8] moyen établi avec chaque PSL sont également conservées quelques soit le jeu de données. Toutefois, ces propriétés environnementales n'expliquent qu'en partie les propriétés de flexibilité des PSLs. En effet, pour une même accessibilité ou un même nombre de contact, les PSLs conservent des flexibilités différentes. Ces résultats confirment l'influence des interactions locales au sein des PSLs sur leur flexibilité et supporte la pertinence de l'utilisation de structures locales pour analyser la flexibilité.

6.6.2 Amélioration de la prédiction de la flexibilité

La prédiction de flexibilité est actuellement testée et améliorée sur le jeu étendu de dynamique moléculaire.

[8] Les contacts à courtes distances avec les 8 résidus encadrant le résidu d'intérêt dans la séquence, ne sont pas pris en compte. Un contact est comptabilisé entre deux résidus si leurs Cα sont à moins de 8 Å.

Des modèles reposant sur les SVMs ou sur les Moindres Carrés Partiels (*Partial Least Square* en anglais, PLS) sont en cours de construction. Des résultats préliminaires ont déjà permis d'obtenir une corrélation de 0,46 entre les B-facteur$_{Norm}$ observés et prédits, soit une augmentation de 0,16 point du coefficient de Pearson par rapport à notre méthode précédente ($r_{Pearson}$ = 0,30, paragraphe 6.5.2.4.2).
Ces modèles seront prochainement évalués sur le jeu de RMN.

Par ailleurs, il convient de souligner que, tout au long de ma thèse, j'ai réalisé des analyses concernant la déformation des structures locales au cours des simulations de dynamique moléculaire. Au niveau méthodologique, différentes représentations et méthodes de mesure et de caractérisation de la déformation ont été testées en prenant en compte les PSLs mais aussi les Blocs Protéiques ou encore les structures secondaires. Ces analyses pourront être utilisées dans le cadre du développement d'une méthode de prédiction de la déformation ou pour approfondir la relation entre la prédiction des structures locales et leur flexibilité.

6.7 Conclusion

Nous avons présenté dans cette étude une approche originale pour étudier et prédire la flexibilité de fragments de séquence au sein des structures protéiques.
L'analyse des propriétés dynamiques associées à chacun des 120 PSLs a conduit à la caractérisation de différents comportements au sein des structures locales. La plus grande flexibilité des boucles par rapport aux structures secondaires répétitives hélicoïdales ou étendues est bien connue. Toutefois, notre analyse décrit plus précisément ce phénomène complexe et permet de mettre en évidence des structures en hélice, étendues ou de connexion plus mobiles que d'autres du même type. Cette description précise de la flexibilité couplée aux spécificités de séquence associées à chaque groupe de PSLs pourrait être d'une aide précieuse dans le domaine du *design* de protéines.
De plus, nos résultats suggèrent que les *erreurs* de prédiction des structures locales (100 % - 63,1 % de prédiction correctes = 36,9 %) ne sont pas seulement dues à de réelles confusions. Ces erreurs sont également dues à l'existence de structures locales flexibles existant plus en tant qu'ensembles conformationnels qu'en tant que structures rigides et prédictibles. Cette vision renforce la pertinence de notre méthode de prédiction des structures locales proposant plusieurs candidats structuraux pour une même séquence cible.

Finalement, nous proposons ici une stratégie de prédiction de la flexibilité à partir de la séquence. Elle est dérivée directement de la prédiction des structures locales. J'ai mis en place une première version du serveur web avec Jean-Christophe Gelly et Alexandre G. de Brevern ; elle est disponible à l'adresse suivante : http://www.dsimb.inserm.fr/dsimb_tools/predyflexy/.

Nos analyses montrent que les résultats de prédictions des structures locales sont suffisamment pertinents pour permettre d'obtenir des résultats déjà compétitifs sans aucun apprentissage supplémentaire. Notre stratégie sépare les zones rigides des zones flexibles de façon très satisfaisante avec peu de confusion. Par ailleurs, un point important de notre travail est la prise en compte de différentes mesures de la flexibilité (B-facteur expérimental et fluctuations observées en simulation de dynamique moléculaire). En effet, la plupart des méthodes de prédiction de la flexibilité reposent uniquement sur les B-facteurs. Or, une étude récente attire l'attention sur l'importance de la contribution aux B-facteurs des fluctuations au sein même du cristal. Cette dernière serait plus importante que la contribution apportée par les fluctuations thermiques des atomes des protéines (Hinsen 2008). Cette étude renforce l'intérêt de prendre en compte plusieurs visions pour obtenir une image réaliste des propriétés dynamiques des protéines. De plus, l'importance du bruit dans les B-facteurs cristallographiques pourrait expliquer les difficultés de la communauté internationale pour le prédire avec un coefficient de corrélation plus élevé que 0,55 (Zhang et al. 2009). Selon, nos résultats, les fluctuations obtenues lors des simulations de dynamique moléculaire sont bien plus prédictibles à partir de la séquence.

7. Conclusion Générale et Perspectives

Tout au long de ma thèse, j'ai travaillé sur la relation entre la séquence, la structure et la flexibilité au sein des protéines.

J'ai eu la chance de collaborer à différentes études m'amenant ainsi à de nombreuses facettes de la bioinformatique structurale :

- la confusion entre les méthodes d'assignation des structures secondaires : *(i)* l'assignation des β-turns (Article 1) et *(ii)* l'impact de cette confusion sur la caractérisation des boucles (Article 2) ont été étudiés.
- la prédiction des boucles *via* l'alphabet structural des blocs protéiques (Article 7).

Les études publiées dans les articles 2 et 7 ont été réalisées dans le cadre de la thèse de Manoj Tyagi effectuée sous la direction de Bernard Offmann (Université de la Réunion).

- la propension des séquences caméléons à adopter des structures non-régulières (Article 3). Ce travail a été effectué par Amine Ghozlane, étudiant en master 1.
- les implications structurales des mutations. L'étude de l'équivalence des acides aminés dans un contexte structurale a permis de proposer un alphabet réduit d'acides aminés (Article 8).
- l'analyse de surfaces protéiques pour l'annotation fonctionnelle grâce au logiciel MED-SUMO (Article 6). Ce travail a été effectué dans le cadre de la thèse d'Olivia Doppelt-Azeroual avec la société MEDIT-SA.
- l'analyse des contacts au sein des protéines. Une analyse systématique de diverses définitions des contacts a été réalisée. De plus, une attention particulière a été portée sur les résultats des méthodes de prédiction du positionnement des chaînes latérales du point de vue des contacts. En fonction de la méthode choisie, seuls 55 à 64 % des contacts sont bien prédits. Ce travail a été réalisé par Guilhem Faure lors de son stage de master 1 sous ma co-direction (Articles 4 et 5).

Par ailleurs, mon principal travail de thèse s'est concentré sur la prédiction des structures locales et de la flexibilité à partir de la séquence. J'ai ainsi tout d'abord pris la suite des travaux réalisés par Cristina Benros durant sa thèse dans le laboratoire. Le Dr. Benros avait développé une librairie de 120 prototypes structuraux représentatifs des structures locales de 11 résidus de long, observées dans les protéines connues. Une méthode de prédiction avait ensuite été mise en place. Mon travail de thèse a permis de développer une nouvelle méthode

de prédiction plus efficace, couplant l'utilisation de données évolutionnaires à une méthode de d'apprentissage élaborée, *i.e.,* les Machines à Vecteurs Supports (SVMs). Un excellent taux de prédiction de 63,1 % a été obtenu, soit un gain de plus de 7 % par rapport à la méthode initiale. Dans un second temps, nous avons étendu notre analyse à l'étude de la flexibilité de structures locales et, réciproquement, de la *prédictibilité* structurale d'une séquence. Deux visions complémentaires de la flexibilité ont été prises en compte : les B-facteurs expérimentaux et les fluctuations des résidus observées lors de simulations de dynamique moléculaire. Ainsi, nous avons montré que les fragments les plus difficiles à prédire sont aussi les plus flexibles. Cette observation suggère que certaines *erreurs* de prédiction seraient dues à une forte flexibilité de certains fragments pour lesquelles l'évaluation en fonction de la seule structure cristallographique n'est pas optimale. Enfin, nous nous sommes appuyés sur l'*informativité* de la prédiction des structures locales pour mettre au point une méthode compétitive de prédiction de la flexibilité en trois classes. Le taux de prédiction de 50,9 % est stable quelque soit la flexibilité de la séquence cible et bien équilibré pour les 3 classes. De plus, point essentiel, une très faible confusion existe entre les régions rigides et flexibles. Cette méthode est actuellement la seule à disposition permettant une prédiction en plus de 2 classes.

Ce travail méthodologique ouvre de nombreuses perspectives dans le domaine de la caractérisation et de la prédiction des structures protéiques.

L'ensemble de mes développements portant sur la prédiction des structures locales et de la flexibilité seront notamment incorporés au travail d'Agnel Praveen Joseph, doctorant d'Alexandre de Brevern. En effet, ce dernier développe actuellement une méthode de reconnaissance de repliements reposant sur les Blocs Protéiques. Les extensions futures de son approche incluront mes méthodes de prédiction.

L'objectif principal, à plus long terme, est de tendre vers la mise en place d'une méthode de prédiction *de novo* des structures protéiques globales. Dans ce cadre, la stratégie actuellement en développement nécessite l'intégration de 3 types de prédictions à partir de la séquence : *(i)* la prédiction des structures locales, *(ii)* la prédiction des contacts entre résidus au sein des structures et *(iii)* la prédiction de la flexibilité.

La prédiction des structures locales et des contacts entre résidus à partir de la séquence permettra d'obtenir un jeu de contraintes locales (à courte distance) et globales (à plus longue distance). A partir de la carte de contact prédite, il sera possible de positionner les carbones α dans l'espace grâce à une méthode développée au laboratoire et reposant sur le couplage d'une méthode des plus courts chemins et d'une minimisation classique. Un affinement crucial sera alors la génération complète du squelette polypeptidique. Les informations

fournies par la prédiction des structures locales seront alors essentiel. Dans ce contexte, la connaissance des règles *grammaticales* guidant les transitions préférentielles entre PSLs au sein des structures sera également particulièrement utile (voir Annexe 2). De même, les indices de confiance comme celui développé pour la prédiction des structures locales seront précieux pour donner plus ou moins de poids à certaines contraintes. Dans une optique similaire à ce dernier point, le 3ème axe de recherche à intégrer sera la prédiction de la flexibilité des structures protéiques à partir de la séquence. Cette prédiction sera également essentielle pour renforcer ou relâcher certaines contraintes au sein de modèles protéiques. Les méthodes *de novo* actuelles les plus performantes, comme ROSETTA, demandent d'énormes puissances de calcul pour la génération de milliers de modèles possibles. Ces modèles sont ensuite triés et sélectionnés. Le projet définit ici vise à intégrer l'utilisation de plus de contraintes dès les premières étapes pour construire à un nombre limité de modèles qui seront ensuite raffinés.

Mais l'étude des protéines ordonnées n'est pas le seul champ d'application possible. Il pourrait être intéressant tester et d'adapter nos approches au domaine de la caractérisation des protéines *désordonnées*. En effet, les régions désordonnées des protéines existent en tant qu'ensemble conformationnel dynamique sans état d'équilibre. Toutefois, ces régions peuvent être caractérisées par une distribution spécifique des angles Φ–Ψ pour chaque résidu. Ainsi, des structures locales transitoires préférentielles peuvent exister en fonction de spécificités de séquence et/ou d'interactions à courte ou longue distance (Ward et al. 2004; Bernado et al. 2005; Receveur-Brechot et al. 2006; Bernado et al. 2007; Mukrasch et al. 2007). Les Blocs Protéiques ou les PSLs pourraient être utilisés pour aider à la caractérisation de ces structures locales transitoires et être intégré à des méthodes d'analyses développées actuellement. Par exemple, afin de caractériser l'échantillonnage conformationnel existant au sein des protéines désordonnées, Bernado et collaborateurs proposent l'algorithme Flexible-Meccano reposant sur la propension des acides aminés pour des angles Φ–Ψ et le volume de leur chaîne latérale (Bernado et al. 2005). Pour créer un conformère, chaque paire d'angles Φ–Ψ de la chaîne polypeptidique est extrait aléatoirement à partir des préférences des résidus dans une banque de boucles résolues par cristallographie. Un grand nombre de conformères sont créés pour prédire un ensemble conformationnel. Cet ensemble généré *in silico* semble être en accord avec des données issues d'expériences de RDC (*Residual Dipolar Coupling*) et SAXS (*Small-Angle X-ray Scattering*) sur la protéine que les auteurs ont utilisée comme exemple. Récemment, une optimisation de cet algorithme pour prendre en compte des interactions à longue distance a été publiée (Bernado et al. 2007; Mukrasch et al. 2007). Notre méthode de

prédiction des structures locales propose un nombre limité de candidats structuraux pour un fragment de séquence cible donné. Ainsi, notre approche pourrait être intégrée à un algorithme d'échantillonnage du type de celui développé par Bernado *et al.* La prédiction d'ensembles conformationnels deviendrait alors moins couteuse en temps de calcul et des interactions à plus longues distances seraient prises en compte grâce à la longueur importante des PSLs de 11 résidus.

De même, cette caractérisation structurale des protéines désordonnées devrait aussi être intéressante pour l'identification de zones particulières importantes pour la reconnaissance moléculaire (*Molecular Recognition Features* en anglais, MoRFs). Les régions MoRFs ont été définies comme des régions transitant du désordre à une structure ordonnée lors d'une liaison avec un partenaire. Elles sont classées en 3 catégories en fonction du type de structure secondaire qu'elles adoptent : α-MoRFs, β-MoRFs and ι-MoRFs (respectivement hélices α, brin β ou boucle). Ce concept est associé à l'idée de l'existence d'une organisation transitoire dans ces régions pouvant apparaître en solution et pouvant ainsi jouer le rôle de sites de contacts primaire avant la liaison avec un partenaire (Fuxreiter et al. 2004; Csizmok et al. 2005; Mohan et al. 2006). Ainsi, l'utilisation de notre prédiction des structures locales pourraient permettre d'identifier et de caractériser plus spécifiquement ces régions dont une meilleure connaissance pourraient être utile pour le développement de nouveaux médicaments (Cheng et al. 2006).

LISTE ET RÉSUMÉS DES PUBLICATIONS

Article 1: Protein beta-turn assignments.

Bornot A., de Brevern A.G.

A classical way to analyze protein 3D structures or models is to investigate their secondary structures. Their predictions are also widely used as a help to build new 3D models. Thus, hundreds of prediction methods have been proposed. Nonetheless before predicting, secondary structure assignment is required even if not trivial. Therefore numerous but diverging assignment methods have been developed. Beta-turns constitute the third most important secondary structures. However, no analysis to compare the beta-turn distributions according to different secondary structure assignment methods has ever been done. We propose in this paper to analyze and evaluate the results of such a comparison. We highlight some important divergence that could have important consequence for the analysis and prediction of beta-turns.

Bioinformation, 2006, 1(5):153-5.

Article 2: Analysis of loop boundaries using different local structure assignment methods.

Tyagi M., Bornot A., Offmann B., de Brevern A.G.

Loops connect regular secondary structures. In many instances, they are known to play important biological roles. Analysis and prediction of loop conformations depend directly on the definition of repetitive structures. Nonetheless, the secondary structure assignment methods (SSAMs) often lead to divergent assignments. In this study, we analyzed, both structure and sequence point of views, how the divergence between different SSAMs affect boundary definitions of loops connecting regular secondary structures. The analysis of SSAMs underlines that no clear consensus between the different SSAMs can be easily found. Because these latter greatly influence the loop boundary definitions, important variations are indeed observed, that is, capping positions are shifted between different SSAMs. On the other hand, our results show that the sequence information in these capping regions are more stable than expected, and, classical and equivalent sequence patterns were found for most of the SSAMs. This is, to our knowledge, the most exhaustive survey in this field as *(i)* various databank have been used leading to similar results without implication of protein redundancy and *(ii)* the first time various SSAMs have been used. This work hence gives new insights into the difficult question of assignment of repetitive structures and addresses the issue of loop boundaries definition. Although SSAMs give very different local structure assignments capping sequence patterns remain efficiently stable.

Prot. Sci,.2009, in press.

Article 3: Protein contacts, inter-residue interactions and side-chain modeling.

Faure G., Bornot A., de Brevern A.G.

Three-dimensional structures of proteins are the support of their biological functions. Their folds are stabilized by contacts between residues. Inner protein contacts are generally described through direct atomic contacts, *i.e.* interactions between side-chain atoms, while contact prediction methods mainly used inter-Calpha distances. In this paper, we have analyzed the protein contacts on a recent high quality non-redundant databank using different criteria. First, we have studied the average number of contacts depending on the distance threshold to define a contact. Preferential contacts between types of amino acids have been highlighted. Detailed analyses have been done concerning the proximity of contacts in the sequence, the size of the proteins and fold classes. The strongest differences have been extracted, highlighting important residues. Then, we studied the influence of five different side-chain conformation prediction methods (SCWRL, IRECS, SCAP, SCATD and SCCOMP) on the distribution of contacts. The prediction rates of these different methods are quite similar. However, using a distance criterion between side chains, the results are quite different, *e.g.* SCAP predicts 50% more contacts than observed, unlike other methods that predict fewer contacts than observed. Contacts deduced are quite distinct from one method to another with at most 75% contacts in common. Moreover, distributions of amino acid preferential contacts present unexpected behaviours distinct from previously observed in the X-ray structures, especially at the surface of proteins. For instance, the interactions involving Tryptophan greatly decrease.

Biochimie, 2008, 90(4):626-39.

Article 4: Analysis of protein contacts into Protein Units.

Faure G.*, Bornot A.*, de Brevern A.G.

Three-dimensional structures of proteins are the support of their biological functions. Their folds are maintained by inter-residue interactions which are one of the main focuses to understand the mechanisms of protein folding and stability. Furthermore, protein structures can be composed of single or multiple functional domains that can fold and function independently. Hence, dividing a protein into domains is useful for obtaining an accurate structure and function determination. In previous studies, we enlightened protein contact properties according to different definitions and developed a novel methodology named Protein Peeling. Within protein structures, Protein Peeling characterizes small successive compact units along the sequence called protein units (PUs). The cutting done by Protein Peeling maximizes the number of contacts within the PUs and minimizes the number of contacts between them. This method is so a relevant tool in the context of the protein folding research and particularly regarding the hierarchical model proposed by George Rose. Here, we accurately analyze the PUs at different levels of cutting, using a non-redundant protein databank. Distribution of PU sizes, number of PUs or their accessibility are screened to determine their common and different features. Moreover, we highlight the preferential amino acid interactions inside and between PUs. Our results show that PUs are clearly an intermediate level between secondary structures and protein structural domains.

**Both authors contributed equally to this work.*
Biochimie, 2009, 91(7): 876-887.

Article 5: Analysis of protein chameleon sequence characteristics.

Ghozlane A., Joseph A.P., Bornot A., de Brevern A.G.

Conversion of local structural state of a protein from an alpha-helix to a beta-strand is usually associated with a major change in the tertiary structure. Similar changes were observed during the self assembly of amyloidogenic proteins to form fibrils, which are implicated in severe diseases conditions, *e.g.*, Alzheimer disease. Studies have emphasized that certain protein sequence fragments known as chameleon sequences do not have a strong preference for either helical or the extended conformations. Surprisingly, the information on the local sequence neighborhood can be used to predict their secondary at a high accuracy level. Here we report a large scale-analysis of chameleon sequences to estimate their propensities to be associated with different local structural states such as alpha-helices, beta-strands and coils. With the help of the propensity information derived from the amino acid composition, we underline their complexity, as more than one quarter of them prefers coil state over to the regular secondary structures. About half of them show preference for both alpha-helix and beta-sheet conformations and either of these two states is favored by the rest.

Bioinformation, 2009, 3(9): 367-369.

Article 6: Functional annotation strategy for protein structures.

Doppelt O., Moriaud F., Bornot A., de Brevern A.G.

Whole-genome sequencing projects are a major source of unknown function proteins. However, as predicting protein function from sequence remains a difficult task, research groups recently started to use 3D protein structures and structural models to bypass it. MED-SuMo compares protein surfaces analyzing the composition and spatial distribution of specific chemical groups (hydrogen bond donor, acceptor, positive, negative, aromatic, hydrophobic, guanidinium, hydroxyl, acyl and glycine). It is able to recognize proteins that have similar binding sites and thus, may perform similar functions. We present here a fine example which points out the interest of MED-SuMo approach for functional structural annotation.

Bioinformation, 2007, 1(9):357-9.

Article 7: Protein short loop prediction in terms of a structural alphabet.

Tyagi M. *, Bornot A.*, Offmann B., de Brevern A.G.

Loops connect regular secondary structures. In many instances, they are known to play crucial biological roles. To bypass the limitation of secondary structure description, we previously defined a structural alphabet composed of 16 structural prototypes, called Protein Blocks (PBs). It leads to an accurate description of every region of 3D protein backbones and has been used in local structure prediction. In the present study, we used our structural alphabet to predict the loops connecting two repetitive structures. Thus, we showed interest to take into account the flanking regions, leading to prediction rate improvement up to 19.8%, but we also underline the sensitivity of such an approach. This research can be used to propose different structures for the loops and to probe and sample their flexibility. It is a useful tool for *ab initio* loop prediction and leads to insights into flexible docking approach.

**Both authors contributed equally to this work.*
Computational Biology and Chemistry, 2009, 33(4):329-33.

Article 8: A reduced amino acid alphabet for understanding and designing protein adaptation to mutation.

Etchebest C., Benros C., Bornot A., Camproux A.C., de Brevern A.G.

Protein sequence world is considerably larger than structure world. In consequence, numerous non-related sequences may adopt similar 3D folds and different kinds of amino acids may thus be found in similar 3D structures. By grouping together the 20 amino acids into a smaller number of representative residues with similar features, sequence world simplification may be achieved. This clustering hence defines a reduced amino acid alphabet (reduced AAA). Numerous works have shown that protein 3D structures are composed of a limited number of building blocks, defining a structural alphabet. We previously identified such an alphabet composed of 16 representative structural motifs (5-residues length) called Protein Blocks (PBs). This alphabet permits to translate the structure (3D) in sequence of PBs (1D). Based on these two concepts, reduced AAA and PBs, we analyzed the distributions of the different kinds of amino acids and their equivalences in the structural context. Different reduced sets were considered. Recurrent amino acid associations were found in all the local structures while other were specific of some local structures (PBs) (e.g Cysteine, Histidine, Threonine and Serine for the alpha-helix Ncap). Some similar associations are found in other reduced AAAs, *e.g.* Ile with Val, or hydrophobic aromatic residues Trp with Phe and Tyr. We put into evidence interesting alternative associations. This highlights the dependence on the information considered (sequence or structure). This approach, equivalent to a substitution matrix, could be useful for designing protein sequence with different features (for instance adaptation to environment) while preserving mainly the 3D fold.

Eur Biophys J., 2007, 36(8):1059-69.

Article 9: A new prediction strategy for long local protein structures using an original description.

Bornot A., Etchebest C., de Brevern A.G.

A relevant and accurate description of three-dimensional (3D) protein structures can be achieved by characterizing recurrent local structures. In a previous study, we developed a library of 120 3D structural prototypes encompassing all known 11-residues long local protein structures and ensuring a good quality of structural approximation. A local structure prediction method was also proposed. Here, overlapping properties of local protein structures in global ones are taken into account to characterize frequent local networks. At the same time, we propose a new long local structure prediction strategy which involves the use of evolutionary information coupled with Support Vector Machines (SVMs). Our prediction is evaluated by a stringent geometrical assessment. Every local structure prediction with a Calpha RMSD less than 2.5 A from the true local structure is considered as correct. A global prediction rate of 63.1% is then reached, corresponding to an improvement of 7.7 points compared with the previous strategy. In the same way, the prediction of 88.33% of the 120 structural classes is improved with 8.65% mean gain. 85.33% of proteins have better prediction results with a 9.43% average gain. An analysis of prediction rate per local network also supports the global improvement and gives insights into the potential of our method for predicting super local structures. Moreover, a confidence index for the direct estimation of prediction quality is proposed. Finally, our method is proved to be very competitive with cutting-edge strategies encompassing three categories of local structure predictions.

Proteins, 2009, 76(3): 570-587.

ANNEXE 1 – *RÉSULTATS DÉTAILLÉS DES DIFFÉRENTES MÉTHODES DE PRÉDICTION DES STRUCTURES LOCALES TESTÉES*

Lors du développement de notre méthode de prédiction des structures locales, différentes stratégies de prédiction ont été testées. Cette annexe résume les résultats obtenus pour chacune de ces stratégies.

Les différentes techniques de prédiction présentées sont :

LR_seq_max5 (méthode originale développée par Benros et collaborateurs (Benros et al. 2006), Experts définis par régression logistique, séquence seule, 5 candidats maximum par position soit 4,2 en moyenne).

LR_seq_5 (de même que *LR_seq_max5* mais réévaluée pour un nombre fixe de 5 candidats par position, *LR_seq_5* correspond à *LR_seq* dans le reste du manuscrit).

LR_seq_max7 (de même que *LR_seq_max5* mais réévaluée pour un nombre maximum de 7 candidats par position, soit 5,1 en moyenne).

LR_PSSM (experts définis par régression logistique, description de la séquence à prédire par un PSSM, nombre fixe de 5 candidats).

$SVM^{\gamma\lambda}\_seq$ (experts définis par SVM, optimisation des SVMs sur le couple (γ,λ), séquence seule, 5 candidats, $SVM^{\gamma\lambda}\_seq$ correspond à *SVM_seq* dans le reste du manuscrit).

$SVM^{\gamma C}\_seq$ (experts définis comme pour $SVM^{\gamma\lambda}\_seq$, mais le couple de paramètres optimisés est ici (γ,C)).

$SVM^{\gamma\lambda}\_PSSM$ (experts définis par SVM, optimisation des SVMs sur le couple (γ,λ), PSSM, 5 candidats, $SVM^{\gamma\lambda}\_PSSM$ correspond à *SVM_PSSM* dans le reste du manuscrit).

$SVM^{\gamma C}\_PSSM$ (de même que $SVM^{\gamma\lambda}\_PSSM$ mais le couple de paramètres optimisés est (γ,C)).

Il est important de souligner que le Q_{120} présenté ici (ou *Mean proportion of true positive per prototype*) est bien le Q_{120} moyen, soit la moyenne des taux de prédiction pour chaque classe structurale. Il donne donc une vision des résultats de prédiction légèrement différente du Q_{120} calculé globalement pour tous les fragments et notamment présenté dans le Tableau 9.

Experts definitions			Logistic regression			
Target sequence window representation			*LR_seq_max 5*	*LR_seq_5*	*LR_seq_max7*	*LR_PSSM*
Mean proportion of true positives per prototype (%)			25.86	27.89	29.02	7.93
Prediction rate (%) (Maximum approximation of 2.5 Å)			51.16	55.48	53.57	35.33
Results per secondary structures categories						
H	*Mean proportion of true positives per prototype (%)*		36.46	40.15	39.06	7.73
	Prediction rate (%)	*1.5 Å*	53.13	58.41	54.14	31.80
		2 Å	63.18	69.38	64.29	46.06
		2.5 Å	70.15	76.80	71.29	56.04
E	*Mean proportion of true positives per prototype (%)*		17.97	19.78	21.56	2.02
	Prediction rate (%)	*1.5 Å*	11.95	12.90	13.27	1.50
		2 Å	31.72	34.48	34.70	7.48
		2.5 Å	52.64	57.43	55.52	29.52
C	*Mean proportion of true positives per prototype (%)*		28.80	30.58	31.80	10,00
	Prediction rate (%)	*1.5 Å*	12.87	13.50	13.43	4.10
		2 Å	24.89	26.17	26.28	9.99
		2.5 Å	42.36	45.06	44.81	22.94
Ext	*Mean proportion of true positives per prototype (%)*		20.43	22.19	23.88	7.30
	Prediction rate (%)	*1.5 Å*	7.08	7.50	7.96	2.58
		2 Å	21.31	22.91	23.84	11.91
		2.5 Å	43.33	47.10	46.90	34.30

SVMs				Gains of SVM^{pk}_PSSM over LR_seq_max 5	Gains of SVM^{pk}_PSSM over LR_seq_5	Gains of SVM^{pk}_PSSM over LR_seq_max7
SVM^{pk}_seq	*SVM^{rc}_seq*	*SVM^{pk}_PSSM*	*SVM^{rc}_PSSM*			
26.61	**26.95**	**33.73**	**33.99**	7.87	5.84	4.71
55.54	**56.55**	**63.13**	**63.21**	11.97	7.65	9.56
38.44	**42.09**	**48.20**	**50.93**	11.74	8.05	9.14
57.85	61.19	67.77	69.52	14.64	9.36	13.63
68.09	72.47	77.63	79.69	14.45	8.25	13.34
75.45	**79.32**	**84.6**	**85.77**	14.45	7.80	13.31
24.07	**22.34**	**32.87**	**30.05**	14.90	13.09	11.31
15.11	14.83	18.71	18.43	6.76	5.81	5.44
39.55	39.53	49.05	46.18	17.33	14.57	14.35
63.18	**62.98**	**73.03**	**69.28**	20.39	15.60	17.51
27.74	**27.65**	**34.51**	**33.13**	5.71	3.93	2.71
13.02	13.41	14.72	15.00	1.85	1.22	1.29
25.36	25.74	29.32	29.35	4.43	3.15	3.04
43.85	**44.27**	**49.47**	**49.43**	7.11	4.41	4.66
21.25	**21.50**	**27.23**	**29.59**	6.80	5.04	3.35
6.76	8.05	8.79	9.40	1.71	1.29	0.83
22.77	23.56	28.11	29.16	6.80	5.20	4.27
48.31	**47.57**	**56.3**	**57.06**	12.97	9.20	9.40

ANNEXE 2 – *DES STRUCTURES LOCALES VERS UNE DESCRIPTION DES STRUCTURES PROTÉIQUES GLOBALES*

Pour aller vers la construction de modèles protéiques globaux, une meilleure connaissance des propriétés d'organisation des PSLs au sein des structures 3D pourrait être un avantage précieux.

Dans ce but, nous avons étudié les transitions préférentielles d'un PSL à l'autre au sein des structures protéiques. Des sous-réseaux privilégiés décrivant des transitions fréquentes entre PSLs ont pu être mis en évidence. Ils caractérisent de longs fragments structuraux et peuvent être décrits en termes de structures super-secondaires. Outre l'intérêt de ces analyses pour la caractérisation de structures de tailles de plus en plus importantes, cette nouvelle vision est également intéressante dès à présent pour évaluer notre stratégie de prédiction des structures locales. Ce développement a été publié dans (Bornot et al. 2009).

2.1 Méthode

Un réseau caractérisant l'enchainement des PSLs a été défini en fonction des probabilités de transitions d'un PSL à un autre au sein des structures protéiques (cf. Figure A2-1).

La probabilité de transition $P_{x\to y}$ d'un PSL *x* au PSL *y* suivant a été calculée somme suit :

$$P_{x\to y} = \frac{N_{x\to y}}{N_x}$$

où $N_{x\to y}$ est le nombre de transitions observées de *x* à *y* et N_x le nombre de transition à partir de *x*.

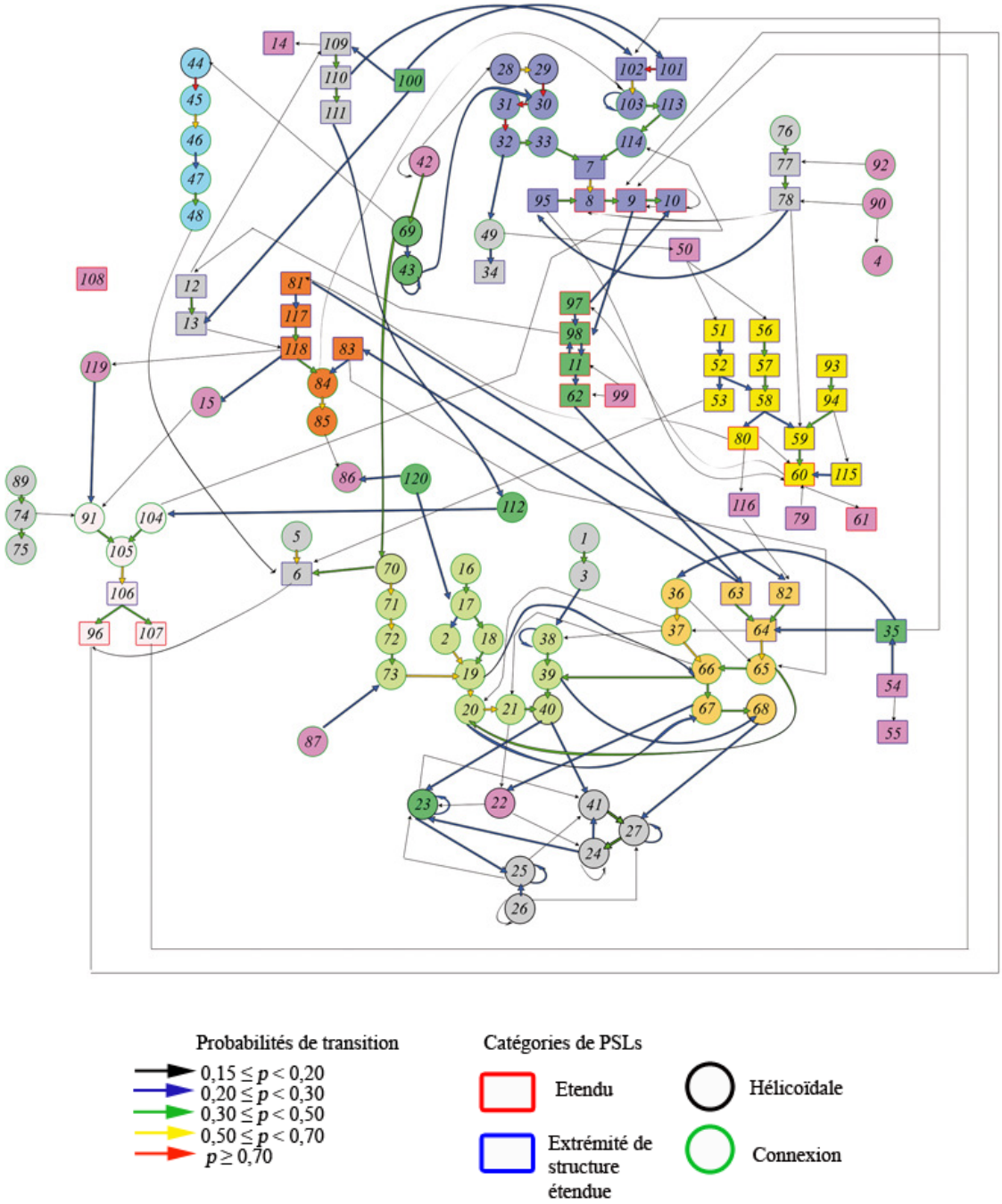

Figure A2-1. Réseau des transitions préférentielles entre PSLs.

Pour des raisons de lisibilité, seules les transitions associées à une probabilité de plus de 0,15 sont illustrées. Ainsi, les transitions indiquées ici ne représentent pas la complexité totale de l'espace des structures protéiques mais il montre en moyenne 45,6 % (σ = 22,2) des transitions d'un PSL donné vers un, deux ou trois autres PSLs. Cette représentation prend donc déjà en compte 51 % des transitions les plus fréquentes observées.

Il est possible de voir que chaque PSL a au moins une probabilité de 0,15 de précéder ou d'être le successeur d'un autre PSL. La seule exception est le PSL 108.

7 sous-réseaux de transitions préférentiels et 3 groupes de PSLs ont été définis (voir paragraphe 2.2 de cette annexe). Chaque sommet (PSL) est coloré en fonction de son appartenance à un sous-réseau ou un groupe donné. Figure extraite de (Bornot et al. 2009).

Ce réseau global a ensuite été itérativement divisé en sous-réseaux en utilisant la procédure suivante :

- une probabilité *p* est initialisée à 0,20. Toutes les probabilités de transitions entre PSLs inférieures à *p* sont alors considérées comme nulles. Ceci induit une première découpe du réseau en *ilôts* ou *sous-réseaux* composés de PSLs transitant les uns vers les autres avec des transitions supérieures à *p*=0,20.
- *p* est augmentée progressivement par pas de 0,02.
- A chaque pas, les grands sous-réseaux de plus de 20 résidus sont à nouveau découpés en sous-réseaux plus petits.

Le processus est arrêté lorsqu'il ne reste aucun sous-réseau de plus de 20 PSLs.

A la fin du processus, *p* était égal à 0,40. Ainsi, tous les PSLs constituant les deux derniers sous-réseaux créés, 6 et 7, sont liés par des probabilités de transitions très élevées et supérieures à 0,40. Plus généralement, au sein des différents sous-réseaux définis, 16,8 % des probabilités de transitions considérées (supérieures à 0,20) sont supérieures à 0,50 et 4 % sont supérieures à 0,70. La probabilité de transition maximale caractérise la transition du PSL n° 101 au PSL 102 dans le 5^{e} sous-réseau, elle est égale à 0,86.

Ainsi, ce processus d'élagage itératif a conduit à 7 sous-réseaux suffisamment peuplés et présentant des probabilités de transition interne significatives. Ces sous-réseaux impliquent 68 PSLs. Les 52 PSLs restants sont soit isolés, soit parties intégrantes de petits sous-réseaux de moins de 5 PSLs. Nous les avons finalement rassemblés en 3 groupes en fonction de leurs propriétés de transition : la taille de ces sous-réseaux, valeurs de leur transition internes et de leurs transitions vers des PSLs extérieurs.

Finalement, 10 groupes de PSLs (7 sous-réseaux et 3 groupes) ont émergés. Les 7 sous-réseaux caractérisent des structures super-secondaires particulières et fréquentes au sein des structures protéiques. Les 3 groupes restants rassemblent des PSLs avec des propriétés de transitions communes. Ils permettent ainsi de mettre en avant des rôles spécifiques de certains PSLs au sein de l'architecture des protéines.

2.2 Résultats

Une analyse des transitions structurales entre PSLs au sein des structures protéiques a mené à la définition d'un réseau global de PSLs présenté en Figure 1. Il caractérise l'enchainement des PSLs au sein de l'architecture des protéines. Il est important de noter que la grande majorité des transitions préférentielles s'effectuent entre PSLs successifs dans le modèle de la

Protéine Hybride, *e.g.*, le PSL 70 transite fréquemment vers le 71 qui lui-même transite vers le 72. La méthode de la protéine hybride a donc bien permis un apprentissage de la succession des structures locales au sein de structures globales (paragraphes 3.3.1 et 3.3.2).

Grâce à un processus itératif d'élagage de ce réseau, nous avons défini dix groupes de PSLs caractérisant les transitions préférentielles des structures locales au sein des structures protéiques. La composition et les transitions des catégories les plus significatives, ou sous-réseaux, sont présentés en Figure A2-2. Chaque sous-réseau comprend un nombre limité de *n* PSLs (*n* allant de 5 à 16). En prenant en compte les embranchements potentiels et les probabilités de transitions, il caractérise une super-structure locale composée au plus de *n* PSLs.

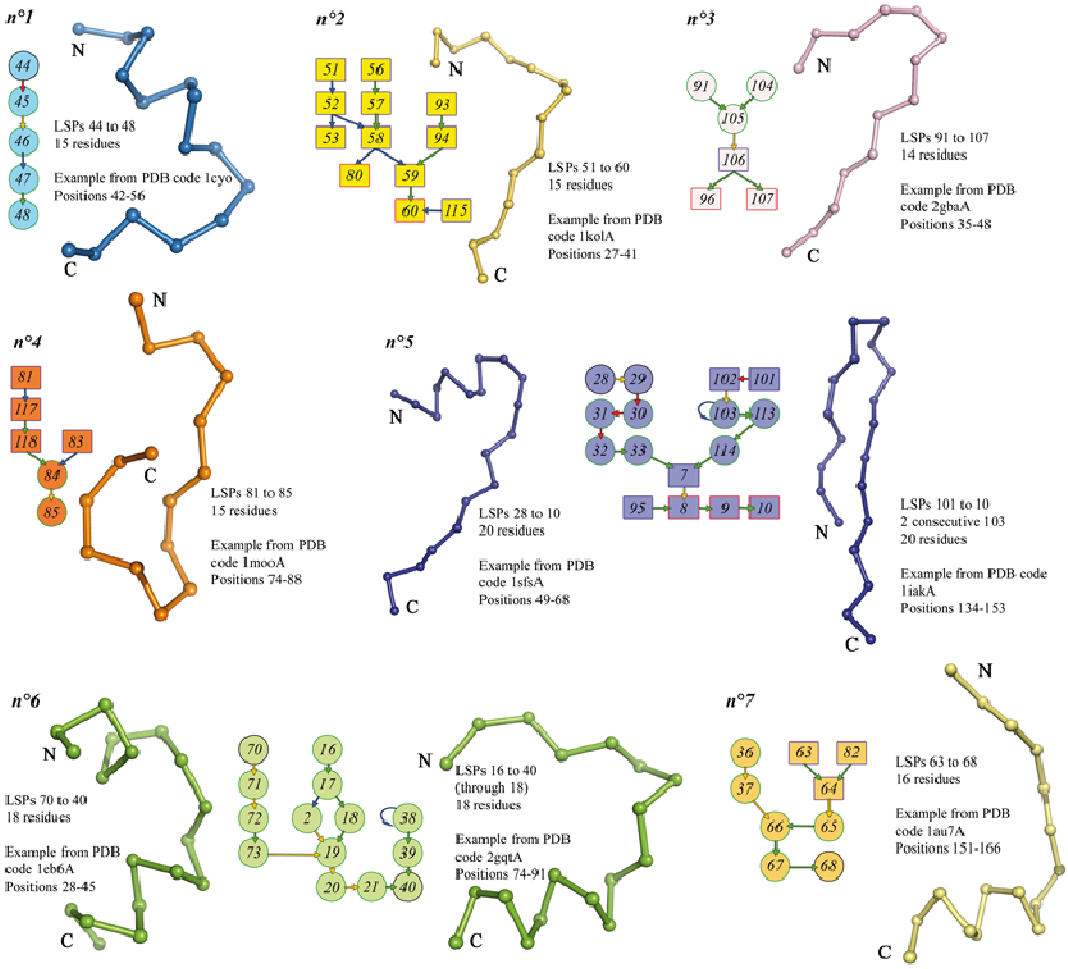

Figure A2-2. Sept sous-réseaux de PSLs présentent des probabilités de transition significatives.

Des exemples de super-structures extraites de protéines sont présentés près des sous-réseaux. Leur position dans les protéines, leur assignation en PSLs et leur longueur sont indiquées. Seules les transitions supérieures à 0,2 sont représentées. Pour la signification des couleurs et des formes des sommets et des arêtes, voir la légende de la Figure A2-1. Figure extraite de (Bornot et al. 2009).

Les sous-réseaux 1 et 2 caractérisent le repliement de séquences pouvant faire jusqu'à 15 résidus de long.

Le sous-réseau 1 est constitué de 5 PSLs. Son point d'entrée est un PSL hélicoïdal suivi de 4 PSLs de connexion spécifiques définissant un β-turn. Ce sous-réseau peut donc être définit comme une structure super-secondaire, *α-C^{cap}-β-turn*, caractérisant une sortie d'hélice. Il est composé de PSLs représentant 2,74 % des fragments. Les PSLs 44, 45 et 46 sont respectivement assignés à 0,67, 0,74 et 0,70 % des fragments et sont ainsi assez fréquents étant donné le nombre important de classes. En effet, chacune des 120 classes est assignée à 0,83 % des fragments en moyenne (σ=0,56, médiane = 0,71 %). En revanche, les PSLs 47 et 48 sont moins fréquents, *i.e.,* ils ne représentent que 0,33 et 0,30 % des fragments. Cette rareté relative renforce la significativité de la succession des PSLs 46 et 47 observée dans les structures protéiques ($P_{46\rightarrow47}$ =27,53 %).

Le sous-réseau 2 est constitué de 12 PSLs représentant 6,57 % des fragments structuraux. A titre d'exemple, la structure super-secondaire présentée en Figure 2 correspond au chemin allant du PSL 51 au PSL 60. Il est composé de quatre types d'extrémités de structures étendues menant à un cœur de structure étendue. Toutes les structures super-secondaires définies par ce sous-réseau commencent par un *β-turn* (coude) suivi d'un brin β court puis d'un second, plus long et presque orthogonal au premier. Nous définiront donc cette structure comme un *turn-ββ-corner*. Son cœur comprend des PSLs assez fréquents. Ainsi, les PSLs 57, 58, 59 et 60 caractérisent un type de coin fréquent regroupant respectivement 0,91, 0,78, 0,63 et 0,79 % des fragments.

Les sous-réseaux 3 et 4 sont tous deux constitués de 6 PSLs. Le sous-réseau 3 propose des chemins alternatifs de quatre PSLs successifs tandis que le sous-réseau 4 caractérise des transitions de 5 PSLs successifs au plus, *i.e.,* des séquences comprenant jusqu'à 14 et 15 résidus de long.

Le sous-réseau 3 présente des voies alternatives pour entrer dans un brin β après un changement de direction induit par un α-turn. Nous appellerons ce sous-réseau *α-turn-β-strand*. Les PSLs de ce sous-réseau représentent 4,05 % des fragments. Les n°106 et 96 sont parmi les 35% de PSLs les plus fréquents, *i.e.,* 0,89 et 0,91 % respectivement.

A l'inverse, le sous-réseau 4 propose différentes terminaison de brin β menant à un changement de direction du squelette polypeptidique du à un *β-turn*. En général, ce changement de direction mène à une autre structure étendue. Néanmoins, dans chaque cas, ce chemin inclut une irrégularité de type β-bulge. Ainsi, le sous-réseau 4 a été identifié en tant

qu'*Irregular β-hairpin-turn*. A nouveau, les PSLs 118, 84 et 85 constituants le cœur de cette super-structure, sont parmi les plus fréquents. Ils représentent respectivement 0,73, 0,82 et 1,81 % des fragments. Le pourcentage total de fragments caractérisés par les PSLs de ce sous-réseau est de 4,24 %.

Les sous-réseaux 5 et 6 sont les plus longs (jusqu'à 20 résidus) et les plus complexes des super-structures définies ici.

Le sous-réseau 5 est constitué de 16 PSLs représentant 19,95 % des fragments. Il décrit deux voies principales pour entrer dans une structure étendues à travers un α-turn. La première voie commence par une structure hélicoïdale tandis que la seconde commence par une série de structures locales étendues. Ce sous-réseau peut être défini comme un *α/β-hairpin*. Les classes structurales le composant sont très peuplées, *i.e.,* 10 d'entre elles représentent plus de un pourcent des fragments chacune.

Le sous-réseau 6 comprend 14 PSLs hélicoïdaux et de connexion regroupant 12,11 % des fragments. Il caractérise trois voies d'entrée dans une structure hélicoïdale. La première voie connecte deux structures hélicoïdales avec une boucle et les deux autres correspondent à de longues séries de structures de connexion. De manière générale, cette série succède à une structure étendue. Nous définissons donc le 6ème sous-réseau comme une super-structure *α/β-loop-α*. De même que pour le sous-réseau 5, le sous-réseau 6 est constitué de PSLs très fréquents, *i.e.,* 6 d'entre eux représentent plus d'1 % des fragments chacun.

Enfin, le sous-réseau 7 comprenant 9 PSLs (jusqu'à 16 résidus successifs), présente des voies alternatives pour entrer dans un structure hélicoïdale, soit exclusivement à travers des PSLs de connexion, soit également à travers des extrémités de structures étendues. L'orthogonalité des structures de connexion ou d'extrémité par rapport à l'hélice nous a conduits à nommer ce sous-réseau un *βα-corner*. Cette structure regroupe des PSLs caractérisant 7,14 % des fragments. Les PSLs 64, 65, 66, 67 et 68 représentent 0,86, 1,02, 1,19, 0,79 et 0,97 % des fragments respectivement.

En fonction de leurs propriétés de transition, les PSLs restants ont été groupés en 3 catégories nommées groupes 8, 9 et 10 (cf. Figure A2-3). Ces derniers sont respectivement composés de 19, 22 et 11 PSLs.

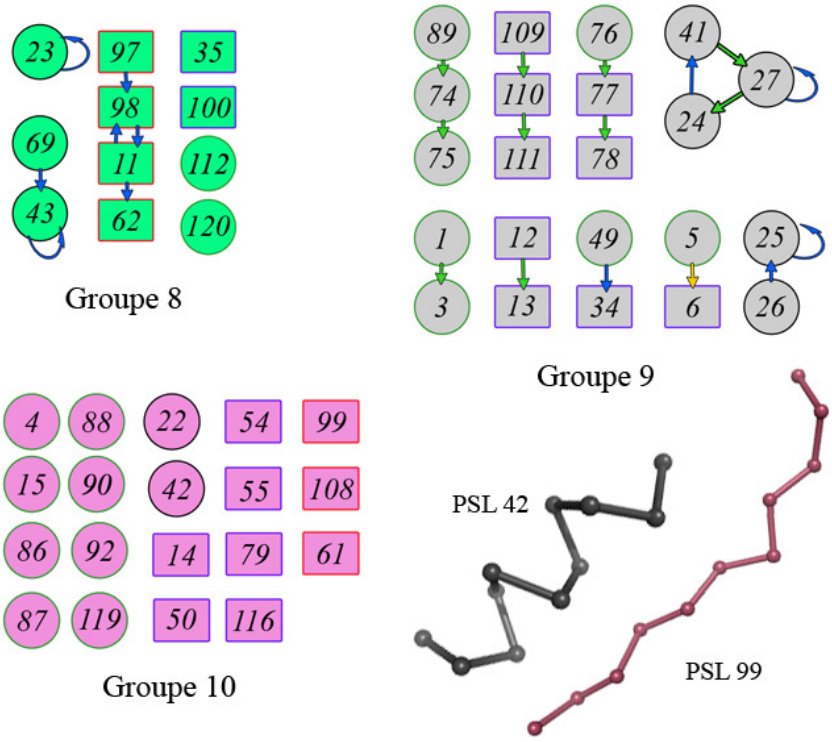

Figure A2-3. Trois groupes de PSLs supplémentaires présentent des propriétés de transition particulières.

Au sein du groupe 8, le processus d'élagage a conduit à la séparation des PSLs 97, 62 et du sous-réseau 98-11. Le PSL 63 a également été séparé du PSL 43. Seules les transitions supérieures à 0,2 sont représentées. Pour la signification des couleurs et des formes des sommets et des arêtes, voir la légende de la Figure A2-1. Les PSLs 42 et 99 sont représentés à titre d'exemple. Figure adaptée de l'annexe 4 de (Bornot et al. 2009).

Le groupe 8 est compose de 10 petits sous-réseaux constitués d'un ou deux sommets, soit 11 PSLs. La principale caractéristique des PSLs constituant ce groupe est qu'ils ont été séparés à la fin du processus d'élagage. Tous présentent de fortes probabilités de transition avec au moins un autre PSL du groupe ou impliqué dans l'un des 7 sous-réseaux décrits plus haut. La forte probabilité de transition du PSL 69 au 43 est par exemple un cas intéressant. En effet, le PSL 43 transite fréquemment vers le PSL 30 du sous-réseau 5 (p = 0,28). De même, la transition du PSL 69 vers le PSL 70 (sous-réseau 6) est fréquente (p = 0,34). Ainsi, les PSLs 69-43 pourrait avoir un rôle d'échangeur au sein de l'architecture des protéines. Le passage par la sortie d'hélice 69 permet d'avoir le choix de se diriger vers un brin β *via* un *α/β-hairpin* (sous-réseau 5) ou vers une hélice α *via* un *α/β-loop-α* (sous-réseau 6).

Le groupe 9 est constitué de 9 petits sous-réseaux incluant 22 PSLs. Chacun d'entre eux est constitué de 2 ou 3 PSLs et présente de fortes probabilités de transitions internes couplées à des transitions faibles avec les autres PSLs. Le sous-réseau présentant les plus faibles probabilités de transitions internes est le circuit fermé composé des PSLs hélicoïdaux 24, 41

et 27 dans lequel par exemple la probabilité de transition entre les PSLs 24 et 41 est de 0,24. La super-structure hélicoïdale décrite par ce petit sous-réseau fait en moyenne 5 PSLs de long mais peut atteindre une longueur de 26 PSLs grâce aux transitions du PSL 27 sur lui-même. Ces PSLs représentent 6,2 % des fragments structuraux et caractérisent 19 % des résidus.
Enfin, le groupe 10 rassemble 19 PSLs associés à de faibles probabilités de transition vers ou de la part d'autres PSLs. Ils se retrouvent donc rapidement isolés au cours du processus d'élagage. La plupart d'entre eux sont parmi les 25 % de PSLs les plus rares. En revanche, les PSLs 42 et 99, respectivement cœur d'hélice et cœur de brin, font figures d'exceptions (cf. Figure 3). Ils sont parmi les PSLs les plus fréquents. Cependant, ils sont liés à beaucoup d'autres PSLs et semblent tenir le rôle de carrefour vers de très nombreuses possibilités. Ainsi, le PSL 42 par exemple ne transite que vers un seul PSL avec une probabilité supérieure à 0,2 : la sortie d'hélice 69 ($P_{42\rightarrow 69}$ = 0,32). Cependant, 8 PSLs hélicoïdaux effectuent plus de 5 % de leurs transitions vers le 42 et réciproquement, ce dernier transite dans plus vers 5 autres PSLs hélicoïdaux avec une probabilité de plus de 0,05.

Dans le paragraphe 4, la prédiction des structures locales a notamment été analysée selon quatre catégories de PSLs proches des structures secondaires. Ces catégories définies précédemment par Benros et collaborateurs permettent une analyse comparative entre les PSLs et cette description bien connue (Benros et al. 2006). Cependant, comme nous l'avons vu, les structures secondaires ne sont pas une description précise et ne caractérisent par exemple aucunement les spécificités structurales des boucles. Ainsi, pour aller plus loin, nous avons tiré partie des 10 catégories de transitions définies dans le paragraphe précédent pour analyser les résultats de prédiction des structures locales.

En fonction de ces catégories, les taux de prédiction basés sur le critère géométrique vont de 37 % (pour le sous-réseau 5/*Irregular β-hairpin-turn*) à 64 % (sous-réseau 1/*α-C^{cap}-β-turn* et 2/*turn-ββ-corner*) pour les 7 premières catégories (voir Tableau A2-1) et peuvent même atteindre 73 % au sein des trois derniers groupes. Les gains par rapport à des prédictions aléatoire ou naïve (par similarité de séquence) sont assez importants : Ils vont de 25,5 % (sous-réseau 4/*Irregular β-hairpin-turn*) à 42,8 % (sous-réseau 5/*α/β-hairpin*) par rapport à une prédiction aléatoire ; et de 8,8 % (groupe 9) à 29,1 % (sous-réseau 2/*turn-ββ-corner*) par rapport à une prédiction *naïve*. Finalement, des gains significatifs de 7,6 % en moyenne sont également observés en comparant la nouvelle stratégie *SVM_PSSM* à la stratégie initiale

(*LR_seq*). Le gain le plus important est observé pour le sous-réseau 2 (16 %) et les plus faibles sont associés aux sous-réseaux 4 et 5 (4,2 et 4,9 % respectivement). Toutes les structures super-secondaires ainsi que tous les groupes de transitions semblent ainsi globalement mieux prédits.

Tableau A2-1. Prédiction des structures locales par catégories de transition.

Analysis of the structural prediction results per transition categories

		1	2	3	4	5	6	7	8	9	10
SVM_PSSM	Proportion of true positives	45.64	29.46	33.24	23.48	45.62	42.42	39.96	37.11	41.09	27.72
	Prediction rate*	64.22	64.22	62.97	36.87	60.88	55.95	59.15	71.71	72.71	57.57
Prediction rate gains over random*		41.83	33.85	38.50	25.53	42.82	38.22	40.60	38.71	37.73	31.76
Prediction rate gains over similar sequences search*		20.81	29.13	24.73	12.33	15.24	15.83	18.95	14.61	8.79	13.37
Prediction rate gains over LR_seq*		6.42	15.98	5.95	4.22	4.90	6.49	6.48	10.97	8.83	6.02

*(approximation < 2.5 Å)

Tableau extrait de (Bornot et al. 2009).

Tableau A2-2. Approximation structurale fournie par la prédiction des structures locales. Détail pour chacun des sous-réseaux.

Average geometrical approximation of the local structure prediction (A)

		All predicted fragments		Fragments correctly predicted according to the geometrical criteria (<2.5 Å)	
		Minimal RMSD[a]	Mean RMSD[a]	Minimal RMSD[a]	Mean RMSD[a]
Transition categories	1	2.22	3.18	1.71	2.82
	2	2.29	3.07	1.91	2.78
	3	2.30	3.13	1.85	2.76
	4	2.80	3.63	2.05	3.04
	5	2.27	3.42	1.60	2.96
	6	2.27	3.19	1.58	2.74
	7	2.17	3.11	1.53	2.68
	8	1.94	2.93	1.43	2.51
	9	1.56	2.43	0.98	1.93
	10	2.17	3.14	1.66	2.66

[a]Over the five candidates per fragment.

Tableau adapté de (Bornot et al. 2009).

Par ailleurs, la précision des prédictions est bien équilibrée en fonction des différents sous-réseaux et groupes de transition (voir Tableau A2-2). Par exemple, pour les 7 premières catégories, tous les fragments prédits sont le sont avec une précision moyenne de 3,07 à 3,67 Å sur les 5 candidats. De plus, dans la liste de ces 5 candidats, le PSL permettant la meilleure approximation est en moyenne à 2,17 Å de la vraie structure locale pour le *βα-corner* ou dans le pire des cas à 2,80 Å pour l'*Irregular β-hairpin-turn*. Les groupes 8 et 9 bénéficient d'une précision encore meilleure avec un RMSD moyen égale à 2,93 et 2,43 Å et un RMSD

minimal (pour le meilleur candidat) tombant à 1,94 et 1,56 Å en moyenne. Ce dernier résultat est lié au fort pourcentage de structures hélicoïdales au sein de ces deux groupes.

Les études précédentes conduites dans le laboratoire avaient permis d'aller de l'identification de *lettres structurales* (BPs) à l'identification de *mots* (PSLs). J'ai présenté ici nos premières analyses permettant d'aller vers une caractérisation de *phrases structurales* récurrentes et de *mots de liaison* particuliers impliquant des directions plus ou moins spécifiques. Une meilleure compréhension de cette *grammaire* permettra de se diriger vers une meilleure connaissance de l'architecture des protéines.

Ainsi, dans cette étude, nous avons analysé les transitions préférentielles des PSLs pour caractériser 7 sous-réseaux décrivant des super-structures récurrentes et 3 groupes de structures locales présentant des propriétés particulières.

Cette description amène de nombreuses nouvelles questions :

- La connaissance de ces transitions préférentielle apporte-t-elle des éléments nouveaux intéressants pour une meilleure compréhension du processus de repliement des protéines ? Plusieurs hypothèses peuvent-être posées. Les éléments carrefours comme le PSL 42 (cœur d'hélice) doivent-il leur rôle à une stabilité intrinsèque importante ? Pourraient-ils être ainsi des noyaux de repliement stabilisant une partie de la protéine pendant que des régions plus sensibles à des interactions longues distances s'organisent ? De même, les sous-réseaux décrivant des transitions préférentielles entre PSL ont-ils une réalité dans le processus de repliement ? La stabilisation d'un des PSLs constituant un sous-réseau donné implique-t-elle préférentiellement la formation de la super-structure locale ? Un croisement de cette analyse avec le résultat du *Protein Peeling* découpant les protéines en sous-unités compactes (paragraphe 2.3.6) pourrait fournir de premiers résultats intéressants.
- De plus, la connaissance de ces sous-réseaux est-elle une information pertinente dans le cadre de la prédiction des structures protéiques globale ? Il serait particulièrement intéressant de raffiner notre stratégie de transition des PSLs en tenant compte de leurs transitions préférentielles.

Une première analyse été effectuée dans cette direction, une évaluation de notre méthode de prédiction *SVM_PSSM* à travers cette description originale des PSLs a été réalisée et donne des résultats prometteurs pour la plupart des sous-réseaux.

RÉFÉRENCES

Akanuma, S., Kigawa, T., and Yokoyama, S. 2002. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proc Natl Acad Sci U S A* **99:** 13549-13553.

Alexandrov, N., and Shindyalov, I. 2003. PDP: protein domain parser. *Bioinformatics* **19:** 429-430.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Andersen, C.A., Palmer, A.G., Brunak, S., and Rost, B. 2002. Continuum secondary structure captures protein flexibility. *Structure* **10:** 175-184.

Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* **36:** D419-425.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181:** 223-230.

Armen, R., Alonso, D.O., and Daggett, V. 2003. The role of alpha-, 3(10)-, and pi-helix in helix-->coil transitions. *Protein Sci* **12:** 1145-1157.

Aurora, R., and Rose, G.D. 1998. Helix capping. *Protein Sci* **7:** 21-38.

Baeten, L., Reumers, J., Tur, V., Stricher, F., Lenaerts, T., Serrano, L., Rousseau, F., and Schymkowitz, J. 2008. Reconstruction of protein backbones from the BriX collection of canonical protein fragments. *PLoS Comput Biol* **4:** e1000083.

Baker, D., and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294:** 93-96.

Banavar, J.R., and Maritan, A. 2007. Physics of proteins. *Annu Rev Biophys Biomol Struct* **36:** 261-280.

Barlow, D.J., and Thornton, J.M. 1988. Helix geometry in proteins. *J Mol Biol* **201:** 601-619.

Baumeister, W., and Steven, A.C. 2000. Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem Sci* **25:** 624-631.

Benhabilès, N., Thomas, A., and Brasseur, R. 2000. Les mécanismes de repliement des protéines solubles. *Biotechnol. Agron. Soc. Environ.* **4 (2):** 71–81.

Benros. 2006. Prediction Proto Cristina.

Benros, C. 2005. Analyse et prediction des structures tridimensionnelles locales des proteines. In *EBGM*, pp. 211. University Paris 7 - Denis Diderot, Paris, France.

Benros, C., de Brevern, A.G., Etchebest, C., and Hazout, S. 2006. Assessing a novel approach for predicting local 3D protein structures from sequence. *Proteins* **62:** 865-880.

Benros, C., De Brevern, A.G., and Hazout, S. 2003. Hybrid Protein Model (HPM): a method for building a library of overlapping local structural prototypes. Sensitivity study and improvements of the training. *IEEE Int Work NNSP* **1:** 53-70.

Benros, C., Martin, J., Tyagi, M., and De Brevern, A.G. 2007. Description of the local protein structure. I. Classical approaches. In *Recent advances in structural bioinformatics*. (ed. A.G. De Brevern). Research Signpost, Trivandrum.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., Di Nola, A., and Haak, J.R. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81:** 3684–3690.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., and Hermans, J. 1981. In Intermolecular Forces. (ed. B. Pullman), pp. 331. D. Reidel Publishing Company: Dordrecht.

Berman, H., Henrick, K., and Nakamura, H. 2003. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10:** 980.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Bernado, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R.W., and Blackledge, M. 2005. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Natl Acad Sci U S A* **102:** 17002-17007.

Bernado, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. 2007. Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* **129:** 5656-5664.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112:** 535-542.

Berry, I.M., Dym, O., Esnouf, R.M., Harlos, K., Meged, R., Perrakis, A., Sussman, J.L., Walter, T.S., Wilson, J., and Messerschmidt, A. 2006. SPINE high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. *Acta Crystallogr D Biol Crystallogr* **62:** 1137-1149.

Bianchet, M.A., Bains, G., Pelosi, P., Pevsner, J., Snyder, S.H., Monaco, H.L., and Amzel, L.M. 1996. The three-dimensional structure of bovine odorant binding protein and its mechanism of odor recognition. *Nat Struct Biol* **3:** 934-939.

Biou, V., Gibrat, J.F., Levin, J.M., Robson, B., and Garnier, J. 1988. Secondary structure prediction: combination of three different methods. *Protein Eng* **2:** 185-191.

Blake, C.C., Koenig, D.F., Mair, G.A., North, A.C., Phillips, D.C., and Sarma, V.R. 1965. Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. *Nature* **206:** 757-761.

Boden, M., and Bailey, T.L. 2006. Identifying sequence regions undergoing conformational change via predicted continuum secondary structure. *Bioinformatics* **22:** 1809-1814.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31:** 365-370.

Boehr, D.D., Dyson, H.J., and Wright, P.E. 2006a. An NMR perspective on enzyme dynamics. *Chem Rev* **106:** 3055-3079.

Boehr, D.D., McElheny, D., Dyson, H.J., and Wright, P.E. 2006b. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313:** 1638-1642.

Boehr, D.D., and Wright, P.E. 2008. Biochemistry. How do proteins interact? *Science* **320:** 1429-1430.

Bolognesi, M., Rosano, C., Losso, R., Borassi, A., Rizzi, M., Wittenberg, J.B., Boffi, A., and Ascenzi, P. 1999. Cyanide binding to Lucina pectinata hemoglobin I and to sperm whale myoglobin: an x-ray crystallographic study. *Biophys J* **77:** 1093-1099.

Bornot, A., and de Brevern, A.G. 2006. Protein beta-turn assignments. *Bioinformation* **1:** 153-155.

Bornot, A., Etchebest, C., and de Brevern, A.G. 2009. A new prediction strategy for long local protein structures using an original description. *Proteins* **76:** 570-587.

Bower, M.J., Cohen, F.E., and Dunbrack, R.L., Jr. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol* **267:** 1268-1282.

Branden, C., and Tooze, J. 1998. *Introduction to Protein Structure*, 2nd ed, New York, pp. 410.

Brautigam, C.A., Sun, S., Piccirilli, J.A., and Steitz, T.A. 1999. Structures of normal single-stranded DNA and deoxyribo-3'-S-phosphorothiolates bound to the 3'-5' exonucleolytic active site of DNA polymerase I from Escherichia coli. *Biochemistry* **38:** 696-704.

Burke, D.F., Deane, C.M., and Blundell, T.L. 2000. Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics* **16:** 513-519.

Butterwick, J.A., Patrick Loria, J., Astrof, N.S., Kroenke, C.D., Cole, R., Rance, M., and Palmer, A.G., 3rd. 2004. Multiple time scale backbone dynamics of homologous thermophilic and mesophilic ribonuclease HI enzymes. *J Mol Biol* **339:** 855-871.

Bystroff, C., and Baker, D. 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* **281:** 565-577.

Bystroff, C., Thorsson, V., and Baker, D. 2000. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* **301:** 173-190.

Camproux, A.C., Gautier, R., and Tuffery, P. 2004. A hidden markov model derived structural alphabet for proteins. *J Mol Biol* **339:** 591-605.

Camproux, A.C., Tuffery, P., Chevrolat, J.P., Boisvieux, J.F., and Hazout, S. 1999. Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng* **12:** 1063-1073.

Canutescu, A.A., Shelenkov, A.A., and Dunbrack, R.L., Jr. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* **12:** 2001-2014.

Cartailler, J.P., and Luecke, H. 2004. Structural and functional characterization of pi bulges and other short intrahelical deformations. *Structure* **12:** 133-144.

Chakrabarti, P., and Bhattacharyya, R. 2007. Geometry of nonbonded interactions involving planar groups in proteins. *Prog Biophys Mol Biol* **95:** 83-137.

Chan, A.W., Hutchinson, E.G., Harris, D., and Thornton, J.M. 1993. Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci* **2:** 1574-1590.

Chandonia, J.M., and Brenner, S.E. 2006. The impact of structural genomics: expectations and outcomes. *Science* **311:** 347-351.

Chang, C.-C., and Lin, C.-J. 2001. LIBSVM: a library for support vector machines. *Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.*

Chellgren, B.W., Miller, A.F., and Creamer, T.P. 2006. Evidence for polyproline II helical structure in short polyglutamine tracts. *J Mol Biol* **361:** 362-371.

Chen, H.F. 2009. Molecular dynamics simulation of phosphorylated KID post-translational modification. *PLoS One* **4:** e6516.

Chen, J.W., Romero, P., Uversky, V.N., and Dunker, A.K. 2006. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res* **5:** 879-887.

Chen, K., Kurgan, L.A., and Ruan, J. 2007a. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct Biol* **7:** 25.

Chen, P., Wang, B., Wong, H.S., and Huang, D.S. 2007b. Prediction of protein B-factors using multi-class bounded SVM. *Protein Pept Lett* **14:** 185-190.

Cheng, Y., LeGall, T., Oldfield, C.J., Mueller, J.P., Van, Y.Y., Romero, P., Cortese, M.S., Uversky, V.N., and Dunker, A.K. 2006. Rational drug design via intrinsically disordered protein. *Trends Biotechnol* **24:** 435-442.

Chou, P.Y., and Fasman, G.D. 1974a. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **13:** 211-222.

Chou, P.Y., and Fasman, G.D. 1974b. Prediction of protein conformation. *Biochemistry* **13:** 222-245.

Clarke, N.D. 1995. Sequence 'minimization': exploring the sequence landscape with simplified sequences. *Curr Opin Biotechnol* **6:** 467-472.

Clore, G.M., and Schwieters, C.D. 2006. Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small alpha/beta protein: a unified picture of high probability, fast atomic motions in proteins. *J Mol Biol* **355:** 879-886.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J.P. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng* **6:** 377-382.

Cowan, P.M., McGavin, S., and North, A.C. 1955. The polypeptide chain configuration of collagen. *Nature* **176:** 1062-1064.

Crooks, G.E., and Brenner, S.E. 2004. Protein secondary structure: entropy, correlations and prediction. *Bioinformatics* **20:** 1603-1611.

Csizmok, V., Bokor, M., Banki, P., Klement, E., Medzihradszky, K.F., Friedrich, P., Tompa, K., and Tompa, P. 2005. Primary contact sites in intrinsically unstructured proteins: the case of calpastatin and microtubule-associated protein 2. *Biochemistry* **44:** 3955-3964.

Cubellis, M.V., Caillez, F., Blundell, T.L., and Lovell, S.C. 2005a. Properties of polyproline II, a secondary structure element implicated in protein-protein interactions. *Proteins* **58:** 880-892.

Cubellis, M.V., Cailliez, F., and Lovell, S.C. 2005b. Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinformatics* **6 Suppl 4:** S8.

Daopin, S., Alber, T., Baase, W.A., Wozniak, J.A., and Matthews, B.W. 1991. Structural and thermodynamic analysis of the packing of two alpha-helices in bacteriophage T4 lysozyme. *J Mol Biol* **221:** 647-667.

de Bakker, P.I., DePristo, M.A., Burke, D.F., and Blundell, T.L. 2003. Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins* **51:** 21-40.

de Brevern, A.G. 2005. New assessment of a structural alphabet. *In Silico Biol* **5:** 283-289.

de Brevern, A.G., Benros, C., Gautier, R., Valadie, H., Hazout, S., and Etchebest, C. 2004. Local backbone structure prediction of proteins. *In Silico Biol* **4:** 381-386.

de Brevern, A.G., Etchebest, C., Benros, C., and Hazout, S. 2007. "Pinning strategy": a novel approach for predicting the backbone structure in terms of protein blocks from sequence. *J Biosci* **32:** 51-70.

de Brevern, A.G., Etchebest, C., and Hazout, S. 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **41:** 271-287.

de Brevern, A.G., and Hazout, S. 2001. Compacting local protein fold with a 'hybrid protein model'. *Theorical Chemistry Accounts* **106:** 36-47.

de Brevern, A.G., and Hazout, S. 2003. 'Hybrid protein model' for optimally defining 3D protein structure fragments. *Bioinformatics* **19:** 345-353.

de Brevern, A.G., Valadie, H., Hazout, S., and Etchebest, C. 2002. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. *Protein Sci* **11:** 2871-2886.

de Brevern, A.G., Wong, H., Tournamille, C., Colin, Y., Le Van Kim, C., and Etchebest, C. 2005. A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). *Biochim Biophys Acta* **1724:** 288-306.

de Groot, B.L., van Aalten, D.M., Scheek, R.M., Amadei, A., Vriend, G., and Berendsen, H.J. 1997. Prediction of protein conformational freedom from distance constraints. *Proteins* **29:** 240-251.

del Sol, A., and Carbonell, P. 2007. The modular organization of domain structures: insights into protein-protein binding. *PLoS Comput Biol* **3:** e239.

DeLano, W.L. 2002. The PyMOL Molecular Graphics System *on World Wide Web* http://www.pymol.org.

Deleage, G., and Roux, B. 1987. An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng* **1:** 289-294.

Dengler, U., Siddiqui, A.S., and Barton, G.J. 2001. Protein structural domains: analysis of the 3Dee domains database. *Proteins* **42:** 332-344.

Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R. 2008. The protein folding problem. *Annu Rev Biophys* **37:** 289-316.

Dobson, C.M. 2003. Protein folding and misfolding. *Nature* **426:** 884-890.

Dodson, G.G., Lane, D.P., and Verma, C.S. 2008. Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO Rep* **9:** 144-150.

Donate, L.E., Rufino, S.D., Canard, L.H., and Blundell, T.L. 1996. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci* **5:** 2600-2616.

Dong. 2008. Stratégie Prediction Dong 2008.

Dong, Q., Wang, X., and Lin, L. 2008. Prediction of protein local structures and folding fragments based on building-block library. *Proteins* **72:** 353-366.

Dong, Q.W., Wang, X.L., and Lin, L. 2007. Methods for optimizing the structure alphabet sequences of proteins. *Comput Biol Med* **37:** 1610-1616.

Doppelt-Azeroual, O. 2009. Développement d'une nouvelle méthode performante de classification des surfaces protéique d'interaction. Optimisations et extensions du logiciel MED-SuMo. In *UFR de science de la vie, INSERM UMR-S 665, DSIMB, MEDIT SA*. Université Paris Diderot - Paris 7, Paris.

Doppelt, O., Moriaud, F., Bornot, A., and De Brevern, A.G. 2007. Functional annotation strategy for protein structures. *Bioinformation* **1:** 357-359.

Dosztanyi, Z., Fiser, A., and Simon, I. 1997. Stabilization centers in proteins: identification, characterization and predictions. *J Mol Biol* **272:** 597-612.

Doucet, N., and Pelletier, J.N. 2007. Simulated annealing exploration of an active-site tyrosine in TEM-1 beta-lactamase suggests the existence of alternate conformations. *Proteins* **69:** 340-348.

Duan, Y., and Kollman, P.A. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282:** 740-744.

Dudev, M., and Lim, C. 2007. Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. *BMC Bioinformatics* **8:** 106.

Dunbrack, R.L., Jr., and Cohen, F.E. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* **6:** 1661-1681.

Dunbrack, R.L., Jr., and Karplus, M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* **230:** 543-574.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., and Uversky, V.N. 2008. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9 Suppl 2:** S1.

Dupuis, F., Sadoc, J.F., Jullien, R., Angelov, B., and Mornon, J.P. 2005. Voro3D: 3D Voronoi tessellations applied to protein structures. *Bioinformatics* **21:** 1715-1716.

Dyson, H.J., and Wright, P.E. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6:** 197-208.

Eddy, S.R. 2004. Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechnol* **22:** 1035-1036.

Edman, P., and Begg, G. 1967. A protein sequenator. *Eur J Biochem* **1:** 80-91.

Efimov, A.V. 1991. Structure of coiled beta-beta-hairpins and beta-beta-corners. *FEBS Lett* **284:** 288-292.

Efimov, A.V. 1993. Patterns of loop regions in proteins. *Curr Opin Struct Biol* **3:** 379-384.

Efimov, A.V. 1994. Common structural motifs in small proteins and domains. *FEBS Lett* **355:** 213-219.

Eisenberg, D. 2003. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc Natl Acad Sci U S A* **100:** 11207-11210.

Eisenmesser, E.Z., Millet, O., Labeikovsky, W., Korzhnev, D.M., Wolf-Watz, M., Bosco, D.A., Skalicky, J.J., Kay, L.E., and Kern, D. 2005. Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438:** 117-121.

Ekman, D., Bjorklund, A.K., Frey-Skott, J., and Elofsson, A. 2005. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol* **348:** 231-243.

Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F.X., Sternberg, M.J., and Oliva, B. 2004. ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res* **32:** D185-188.

Eswar, N., Ramakrishnan, C., and Srinivasan, N. 2003. Stranded in isolation: structural role of isolated extended strands in proteins. *Protein Eng* **16:** 331-339.

Etchebest, C., Benros, C., Bornot, A., Camproux, A.C., and de Brevern, A.G. 2007. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *Eur Biophys J* **36:** 1059-1069.

Etchebest, C., Benros, C., Hazout, S., and de Brevern, A.G. 2005. A structural alphabet for local protein structures: improved prediction methods. *Proteins* **59:** 810-827.

Eyal, E., Najmanovich, R., McConkey, B.J., Edelman, M., and Sobolev, V. 2004. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J Comput Chem* **25:** 712-724.

Faure, G., Bornot, A., and de Brevern, A.G. 2008. Protein contacts, inter-residue interactions and side-chain modelling. *Biochimie* **90:** 626-639.

Faure, G., Bornot, A., and de Brevern, A.G. 2009. Analysis of protein contacts into Protein Units. *Biochimie* **91:** 876-887.

Fawcett, T. 2003. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *Intelligent Enterprise Technologies Laboratory*.

Fernandez-Fuentes, N., Hermoso, A., Espadaler, J., Querol, E., Aviles, F.X., and Oliva, B. 2004. Classification of common functional loops of kinase super-families. *Proteins* **56:** 539-555.

Ferron, F., Longhi, S., Canard, B., and Karlin, D. 2006. A practical overview of protein disorder prediction methods. *Proteins* **65:** 1-14.

Fersht, A.R. 2008. From the first protein structures to our current knowledge of protein folding: delights and scepticisms. *Nat Rev Mol Cell Biol* **9:** 650-654.

Fetrow, J.S., Palumbo, M.J., and Berg, G. 1997. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins* **27:** 249-271.

Fink, A.L. 2005. Natively unfolded proteins. *Curr Opin Struct Biol* **15:** 35-41.

Fiser, A., Do, R.K., and Sali, A. 2000. Modeling of loops in protein structures. *Protein Sci* **9:** 1753-1773.

Fitzkee, N.C., Fleming, P.J., Gong, H., Panasik, N., Jr., Street, T.O., and Rose, G.D. 2005. Are proteins made from a limited parts list? *Trends Biochem Sci* **30:** 73-80.

Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E.Z., Thorpe, M., and Gerstein, M. 2006. The Database of Macromolecular Motions: new features added at the decade mark. *Nucleic Acids Res* **34:** D296-301.

Fodje, M.N., and Al-Karadaghi, S. 2002. Occurrence, conformational features and amino acid propensities for the pi-helix. *Protein Eng* **15:** 353-358.

Fourrier, L., Benros, C., and de Brevern, A.G. 2004. Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics* **5:** 58.

Freedberg, D.I., Ishima, R., Jacob, J., Wang, Y.X., Kustanovich, I., Louis, J.M., and Torchia, D.A. 2002. Rapid structural fluctuations of the free HIV protease flaps in solution: relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Sci* **11:** 221-232.

Friedman, R., Nachliel, E., and Gutman, M. 2006. Fatty acid binding proteins: same structure but different binding mechanisms? Molecular dynamics simulations of intestinal fatty acid binding protein. *Biophys J* **90:** 1535-1545.

Frishman, D., and Argos, P. 1995. Knowledge-based protein secondary structure assignment. *Proteins* **23:** 566-579.

Fuchs, P.F., and Alix, A.J. 2005. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. *Proteins* **59:** 828-839.

Fuchs, P.F., Bonvin, A.M., Bochicchio, B., Pepe, A., Alix, A.J., and Tamburro, A.M. 2006. Kinetics and thermodynamics of type VIII beta-turn formation: a CD, NMR, and microsecond explicit molecular dynamics study of the GDNP tetrapeptide. *Biophys J* **90:** 2745-2759.

Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. 2004. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* **338:** 1015-1026.

Garnier, J., Osguthorpe, D.J., and Robson, B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* **120:** 97-120.

Garzon, J.I., Kovacs, J., Abagyan, R., and Chacon, P. 2007. DFprot: a webtool for predicting local chain deformability. *Bioinformatics* **23:** 901-902.

Gelly, J.C., de Brevern, A.G., and Hazout, S. 2006a. 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments. *Bioinformatics* **22:** 129-133.

Gelly, J.C., Etchebest, C., Hazout, S., and de Brevern, A.G. 2006b. Protein Peeling 2: a web server to convert protein structures into series of protein units. *Nucleic Acids Res* **34:** W75-78.

George, D.G., Barker, W.C., and Hunt, L.T. 1986. The protein identification resource (PIR). *Nucleic Acids Res* **14:** 11-15.

Ghozlane, A., Joseph, A.P., Bornot, A., and De Brevern, A.G. 2009. Analysis of protein chameleon sequence characteristics. *Bioinformation* **3:** 367-369.

Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T., and Ben-Tal, N. 2005. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins* **58:** 610-617.

Greenfield, N.J. 2006. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* **1:** 2876-2890.

Gromiha, M.M. 2009. Multiple contact network is a key determinant to protein folding rates. *J Chem Inf Model* **49:** 1130-1135.

Gromiha, M.M., and Selvaraj, S. 1999. Importance of long-range interactions in protein folding. *Biophys Chem* **77:** 49-68.
Gromiha, M.M., and Selvaraj, S. 2001. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J Mol Biol* **310:** 27-32.
Grossfield, A., Pitman, M.C., Feller, S.E., Soubias, O., and Gawrisch, K. 2008. Internal hydration increases during activation of the G-protein-coupled receptor rhodopsin. *J Mol Biol* **381:** 478-486.
Gu, J., Gribskov, M., and Bourne, P.E. 2006. Wiggle-predicting functionally flexible regions from primary sequence. *PLoS Comput Biol* **2:** e90.
Guo, J.T., Jaromczyk, J.W., and Xu, Y. 2007. Analysis of chameleon sequences and their implications in biological processes. *Proteins* **67:** 548-558.
Guo, J.T., Xu, D., Kim, D., and Xu, Y. 2003. Improving the performance of DomainParser for structural domain partition using neural network. *Nucleic Acids Res* **31:** 944-952.
Guruprasad, K., and Rajkumar, S. 2000. Beta-and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. *J Biosci* **25:** 143-156.
Hagen, J.B. 2000. The origins of bioinformatics. *Nat Rev Genet* **1:** 231-236.
Han, K.F., and Baker, D. 1995. Recurring local sequence motifs in proteins. *J Mol Biol* **251:** 176-187.
Han, K.F., and Baker, D. 1996. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci U S A* **93:** 5814-5818.
Hastie, T., Tibshirani, R., and Friedman, J. 2001. *The Elements of statistical learning – Data mining, Inference, and Prediction*. Springer Series in Statistic.
Hazout, S. 2007. Entropy-derived measures for assessing the accuracy of N-state prediction algorithms. In *Recent Advances in Structural Bioinformatics*. (ed. A.G. De Brevern). Research Signpost, Trivandrum, Kerala, India.
Hecht, M.H., Das, A., Go, A., Bradley, L.H., and Wei, Y. 2004. De novo proteins from designed combinatorial libraries. *Protein Sci* **13:** 1711-1723.
Henzel, W.J., Watanabe, C., and Stults, J.T. 2003. Protein identification: the origins of peptide mass fingerprinting. *J Am Soc Mass Spectrom* **14:** 931-942.
Hess, B., Bekker, H., Berendsen, H.J.C., and Fraaije, J.G.E.M. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **18:** 1463–1472.
Hillig, R.C., Hanzal-Bayer, M., Linari, M., Becker, J., Wittinghofer, A., and Renault, L. 2000. Structural and biochemical properties show ARL3-GDP as a distinct GTP binding protein. *Structure* **8:** 1239-1245.
Hinsen, K. 2008. Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics* **24:** 521-528.
Ho, B.K., and Agard, D.A. 2009. Probing the flexibility of large conformational changes in protein structures through local perturbations. *PLoS Comput Biol* **5:** e1000343.
Ho, B.K., Thomas, A., and Brasseur, R. 2003. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein Sci* **12:** 2508-2522.
Hodsdon, M.E., and Cistola, D.P. 1997. Ligand binding alters the backbone mobility of intestinal fatty acid-binding protein as monitored by 15N NMR relaxation and 1H exchange. *Biochemistry* **36:** 2278-2290.
Holland, T.A., Veretnik, S., Shindyalov, I.N., and Bourne, P.E. 2006. Partitioning protein structures into domains: why is it so difficult? *J Mol Biol* **361:** 562-590.
Holm, L., and Sander, C. 1994. Parser for protein folding units. *Proteins* **19:** 256-268.
Holm, L., and Sander, C. 1998. Dictionary of recurrent domains in protein structures. *Proteins* **33:** 88-96.
Hosseini, S.R., Sadeghi, M., Pezeshk, H., Eslahchi, C., and Habibi, M. 2008. PROSIGN: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C(alpha) atoms. *Comput Biol Chem* **32:** 406-411.
Hsu, C.W., Chang, C.C., and Lin, C.J. 2003. A practical guide to support vector classification. *Tech. Rep., Department of computer science and information engineering, National Taiwan University, Tapei, Taiwan. [Available at http://www.csie.ntu.edu.tw/~cjlin/papers.html]*.

Huang, Y.J., and Montelione, G.T. 2005. Structural biology: proteins flex to function. *Nature* **438:** 36-37.

Humphrey, W., Dalke, A., and Schulten, K. 1996. VMD: visual molecular dynamics. *J Mol Graph* **14:** 33-38, 27-38.

Hunter, C.G., and Subramaniam, S. 2003. Protein fragment clustering and canonical local shapes. *Proteins* **50:** 580-588.

Hutchinson, E.G., and Thornton, J.M. 1996. PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci* **5:** 212-220.

Idiyatullin, D., Daragan, V.A., and Mayo, K.H. 2003. Protein dynamics using frequency-dependent order parameters from analysis of NMR relaxation data. *J Magn Reson* **161:** 118-125.

Ihaka, R., and Gentleman, R. 1996. R: a language for data analysis and graphics. *J. Comp. Graph. Stat.* **5:** 229-314.

Illergard, K., Ardell, D.H., and Elofsson, A. 2009. Structure is three to ten times more conserved than sequence-A study of structural response in protein cores. *Proteins*.

Imai, K., and Mitaku, S. 2005. Mechanisms of secondary structure breakers in soluble proteins. *Biophysics* **1:** 55-65.

Izarzugaza, J.M., Grana, O., Tress, M.L., Valencia, A., and Clarke, N.D. 2007. Assessment of intramolecular contact predictions for CASP7. *Proteins* **69 Suppl 8:** 152-158.

Jacoboni, I., Martelli, P.L., Fariselli, P., Compiani, M., and Casadio, R. 2000. Predictions of protein segments with the same aminoacid sequence and different secondary structure: a benchmark for predictive methods. *Proteins* **41:** 535-544.

Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F. 2001. Protein flexibility predictions using graph theory. *Proteins* **44:** 150-165.

Jambon, M., Imberty, A., Deleage, G., and Geourjon, C. 2003. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **52:** 137-145.

James, L.C., and Tawfik, D.S. 2003. Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. *Trends Biochem Sci* **28:** 361-368.

Jauch, R., Yeo, H.C., Kolatkar, P.R., and Clarke, N.D. 2007. Assessment of CASP7 structure predictions for template free targets. *Proteins* **69 Suppl 8:** 57-67.

Jia, J., Borregaard, N., Lollike, K., and Cygler, M. 2001. Structure of Ca(2+)-loaded human grancalcin. *Acta Crystallogr D Biol Crystallogr* **57:** 1843-1849.

Jin, Y., and Dunbrack, R.L., Jr. 2005. Assessment of disorder predictions in CASP6. *Proteins* **61 Suppl 7:** 167-175.

Joachims, T. 1999. Making large-Scale SVM Learning Practical. In *Advances in Kernel Methods - Support Vector Learning*. (eds. B. Schölkopf, C. Burges, and A. Smola). MIT-Press.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292:** 195-202.

Joshi, R.R. 2007. A Decade of Computing to Traverse the Labyrinth of Protein Domains. *Current Bioinformatics* **2:** 113-131.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577-2637.

Kabsch, W., and Sander, C. 1984. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci U S A* **81:** 1075-1078.

Karchin, R., Cline, M., Mandel-Gutfreund, Y., and Karplus, K. 2003. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* **51:** 504-514.

Karplus, M., and Weaver, D.L. 1976. Protein-folding dynamics. *Nature* **260:** 404-406.

Karplus, M., and Weaver, D.L. 1994. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci* **3:** 650-668.

Karplus, P.A., and Schulz, G.E. 1985. Prediction of chain flexibility in Proteins - A tool for the selection of peptide antigens. *Naturwissenschaften* **72:** 212-213.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181:** 662-666.

King, S.M., and Johnson, W.C. 1999. Assigning secondary structure from protein coordinate data. *Proteins* **35:** 313-320.
Klepeis, J.L., and Floudas, C.A. 2003. ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys J* **85:** 2119-2146.
Kleywegt, G.J., and Jones, T.A. 1997. Model building and refinement practice. *Methods Enzymol* **277:** 208-230.
Koch, O., and Klebe, G. 2009. Turns revisited: a uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins* **74:** 353-367.
Kohonen, T. 1989. An introduction to neural computing. *Neural Networks* **1**.
Kohonen, T. 1997. *Self-organizing maps.*, 2nd ed. Springer-Verlag, Berlin, pp. 376.
Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. 2002. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* **323:** 297-307.
Korzhnev, D.M., and Kay, L.E. 2008. Probing invisible, low-populated States of protein molecules by relaxation dispersion NMR spectroscopy: an application to protein folding. *Acc Chem Res* **41:** 442-451.
Koshland, D.E. 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci U S A* **44:** 98-104.
Kovacs, J.A., Chacon, P., and Abagyan, R. 2004. Predictions of protein flexibility: first-order measures. *Proteins* **56:** 661-668.
Krebs, W.G., and Gerstein, M. 2000. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res* **28:** 1665-1675.
Kroemer, R.T. 2007. Structure-based drug design: docking and scoring. *Curr Protein Pept Sci* **8:** 312-328.
Ku, S.Y., and Hu, Y.J. 2008. Protein structure search and local structure characterization. *BMC Bioinformatics* **9:** 349.
Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. 2004a. Profile-based string kernels for remote homology detection and motif extraction. *Proc IEEE Comput Syst Bioinform Conf***:** 152-160.
Kuang, R., Leslie, C.S., and Yang, A.S. 2004b. Protein backbone angle prediction with machine learning approaches. *Bioinformatics* **20:** 1612-1621.
Kuhlman, B., and Baker, D. 2004. Exploring folding free energy landscapes using computational protein design. *Curr Opin Struct Biol* **14:** 89-95.
Kullback, S., and Leibler, C.S. 1951. On information and sufficiency. *Ann. Math Stat.* **22:** 79-86.
Kumar, S., and Bansal, M. 1998. Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys J* **75:** 1935-1944.
Kwasigroch, J.M., Chomilier, J., and Mornon, J.P. 1996. A global taxonomy of loops in globular proteins. *J Mol Biol* **259:** 855-872.
Labesse, G., Colloc'h, N., Pothier, J., and Mornon, J.P. 1997. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. *Comput Appl Biosci* **13:** 291-295.
Lakomek, N.A., Lange, O.F., Walter, K.F., Fares, C., Egger, D., Lunkenheimer, P., Meiler, J., Grubmuller, H., Becker, S., de Groot, B.L., et al. 2008. Residual dipolar couplings as a tool to study molecular recognition of ubiquitin. *Biochem Soc Trans* **36:** 1433-1437.
Lange, O.F., Lakomek, N.A., Fares, C., Schroder, G.F., Walter, K.F., Becker, S., Meiler, J., Grubmuller, H., Griesinger, C., and de Groot, B.L. 2008. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320:** 1471-1475.
Laskowski, R.A., Chistyakov, V.V., and Thornton, J.M. 2005a. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res* **33:** D266-268.
Laskowski, R.A., Watson, J.D., and Thornton, J.M. 2005b. Protein function prediction using local 3D templates. *J Mol Biol* **351:** 614-626.

Lavery, R., and Sacquin-Mora, S. 2007. Protein mechanics: a route from structure to function. *J Biosci* **32:** 891-898.
Le, Q., Pollastri, G., and Koehl, P. 2009. Structural alphabets for protein structure classification: a comparison study. *J Mol Biol* **387:** 431-450.
Leach, A.R. 2001. *Molecular Modelling. Principles and Applications.*, Pearson Education Limited ed.
Lee, K.H., Benson, D.R., and Kuczera, K. 2000. Transitions from alpha to pi helix observed in molecular dynamics simulations of synthetic peptides. *Biochemistry* **39:** 13737-13747.
Lesk, A.M., and Rose, G.D. 1981. Folding units in globular proteins. *Proc Natl Acad Sci U S A* **78:** 4304-4308.
Leszczynski, J.F., and Rose, G.D. 1986. Loops in globular proteins: a novel category of secondary structure. *Science* **234:** 849-855.
Levin, J.M., and Garnier, J. 1988. Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochim Biophys Acta* **955:** 283-295.
Levinthal, C. 1966. Molecular model-building by computer. *Sci Am* **214:** 42-52.
Levinthal, C. 1968. Are there pathways for protein folding? *J. Chem. Phys.* **65:** p. 44–45.
Levitt, M. 1978. Conformational preferences of amino acids in globular proteins. *Biochemistry* **17:** 4277-4285.
Levitt, M., and Greer, J. 1977. Automatic identification of secondary structure in globular proteins. *J Mol Biol* **114:** 181-239.
Lewis, D.P., Jebara, T., and Noble, W.S. 2006. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinformatics* **22:** 2753-2760.
Li, J., Wang, J., and Wang, W. 2008. Identifying folding nucleus based on residue contact networks of proteins. *Proteins* **71:** 1899-1907.
Li, Q., Zhou, C., and Liu, H. 2009. Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities. *Proteins* **74:** 820-836.
Li, T., Fan, K., Wang, J., and Wang, W. 2003. Reduction of protein sequence complexity by residue grouping. *Protein Eng* **16:** 323-330.
Li, W., Liang, S., Wang, R., Lai, L., and Han, Y. 1999a. Exploring the conformational diversity of loops on conserved frameworks. *Protein Eng* **12:** 1075-1086.
Li, W., Liu, Z., and Lai, L. 1999b. Protein loops on structurally similar scaffolds: database and conformational analysis. *Biopolymers* **49:** 481-495.
Lindahl, E., Hess, B., and van der Spoel, D. 2001. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.* **7:** 306-317.
Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. 2003. Protein disorder prediction: implications for structural proteomics. *Structure* **11:** 1453-1459.
Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N.C. 2008. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **36:** D475-479.
Lipari, G., and Szabo, A. 1982. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. I. Theory and range of validity. *JACS* **104:** 4546–4559.
Liu, J., Lou, Y., Yokota, H., Adams, P.D., Kim, R., and Kim, S.H. 2005. Crystal structure of a PhoU protein homologue: a new class of metalloprotein containing multinuclear iron clusters. *J Biol Chem* **280:** 15960-15966.
Liu, X., Zhang, L.M., Guan, S., and Zheng, W.M. 2003. Distances and classification of amino acids for different protein secondary structures. *Phys Rev E Stat Nonlin Soft Matter Phys* **67:** 051927.
Llorca, O., Martin-Benito, J., Grantham, J., Ritco-Vonsovici, M., Willison, K.R., Carrascosa, J.L., and Valpuesta, J.M. 2001. The 'sequential allosteric ring' mechanism in the eukaryotic chaperonin-assisted folding of actin and tubulin. *EMBO J* **20:** 4065-4075.
Low, B.W., and Baybutt, R.B. 1952. The pi-helix-A hydrogen bonded configuration of the polypeptide chain. *J Am Chem Soc* **74:** 5806.

Majumdar, I., Krishna, S.S., and Grishin, N.V. 2005. PALSSE: a program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics* **6:** 202.

Malkov, S.N., Zivkovic, M.V., Beljanski, M.V., Stojanovic, S.D., and Zaric, S.D. 2009. A reexamination of correlations of amino acids with particular secondary structures. *Protein J* **28:** 74-86.

Mamonova, T., Hespenheide, B., Straub, R., Thorpe, M.F., and Kurnikova, M. 2005. Protein flexibility using constraints from molecular dynamics simulations. *Phys Biol* **2:** S137-147.

Markley, J.L., Ulrich, E.L., Berman, H.M., Henrick, K., Nakamura, H., and Akutsu, H. 2008. BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR* **40:** 153-155.

Martin, J., Letellier, G., Marin, A., Taly, J.F., de Brevern, A.G., and Gibrat, J.F. 2005. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct Biol* **5:** 17.

Matthews, B.W. 1976. X-ray crystallographic studies of proteins. *Annu. Rev. Phys. Chem.* **27:** 493–523.

Matthews, B.W. 2007. Protein Structure Initiative: getting into gear. *Nat Struct Mol Biol* **14:** 459-460.

Mezei, M. 1998. Chameleon sequences in the PDB. *Protein Eng* **11:** 411-414.

Micheletti, C., Seno, F., and Maritan, A. 2000. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins* **40:** 662-674.

Miick, S.M., Casteel, K.M., and Millhauser, G.L. 1993. Experimental molecular dynamics of an alanine-based helical peptide determined by spin label electron spin resonance. *Biochemistry* **32:** 8014-8021.

Millhauser, G.L. 1995. Views of helical peptides: a proposal for the position of 3(10)-helix along the thermodynamic folding pathway. *Biochemistry* **34:** 3873-3877.

Milner-White, E.J. 1990. Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites. *J Mol Biol* **216:** 386-397.

Minor, D.L., Jr., and Kim, P.S. 1996. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380:** 730-734.

Mitchell, J.B., Thornton, J.M., Singh, J., and Price, S.L. 1992. Towards an understanding of the arginine-aspartate interaction. *J Mol Biol* **226:** 251-262.

Miyazawa, S., and Jernigan, R.L. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256:** 623-644.

Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K., and Uversky, V.N. 2006. Analysis of molecular recognition features (MoRFs). *J Mol Biol* **362:** 1043-1059.

Moore, A.D., Bjorklund, A.K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. 2008. Arrangements in the modular evolution of proteins. *Trends Biochem Sci* **33:** 444-451.

Moreau, V., Fleury, C., Piquer, D., Nguyen, C., Novali, N., Villard, S., Laune, D., Granier, C., and Molina, F. 2008. PEPOP: computational design of immunogenic peptides. *BMC Bioinformatics* **9:** 71.

Moreau, V., Granier, C., Villard, S., Laune, D., and Molina, F. 2006. Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics* **22:** 1088-1095.

Moult, J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* **15:** 285-289.

Mukrasch, M.D., Markwick, P., Biernat, J., Bergen, M., Bernado, P., Griesinger, C., Mandelkow, E., Zweckstetter, M., and Blackledge, M. 2007. Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *J Am Chem Soc* **129:** 5235-5243.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247:** 536-540.

Natalello, A., Ami, D., Brocca, S., Lotti, M., and Doglia, S.M. 2005. Secondary structure, conformational stability and glycosylation of a recombinant Candida rugosa lipase studied by Fourier-transform infrared spectroscopy. *Biochem J* **385:** 511-517.

Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A., Nabuurs, S.B., Guntert, P., Livny, M., Markley, J.L., Nilges, M., et al. 2005. RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins* **59:** 662-672.

Noguchi, T., Matsuda, H., and Akiyama, Y. 2001. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res* **29:** 219-220.

Nolting, B., and Andert, K. 2000. Mechanism of protein folding. *Proteins* **41:** 288-298.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A.K. 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61 Suppl 7:** 176-182.

Offmann, B., Tyagi, M., and de Brevern, A.G. 2007. Local Protein Structures. *Current Bioinformatics* **2:** 165-202.

Okazaki, K., and Takada, S. 2008. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc Natl Acad Sci U S A* **105:** 11182-11187.

Oliva, B., Bates, P.A., Querol, E., Aviles, F.X., and Sternberg, M.J. 1997. An automated classification of the structure of protein loops. *J Mol Biol* **266:** 814-830.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH--a hierarchic classification of protein domain structures. *Structure* **5:** 1093-1108.

Pal, L., Chakrabarti, P., and Basu, G. 2003. Sequence and structure patterns in proteins from an analysis of the shortest helices: implications for helix nucleation. *J Mol Biol* **326:** 273-291.

Pauling, L., and Corey, R.B. 1951. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A* **37:** 251-256.

Pauling, L., Corey, R.B., and Branson, H.R. 1951. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37:** 205-211.

Pauling, L., and Niemann, C. 1939. The Structure of Proteins. *JACS* **61**.

Pavone, V., Gaeta, G., Lombardi, A., Nastri, F., Maglio, O., Isernia, C., and Saviano, M. 1996. Discovering protein secondary structures: classification and description of isolated alpha-turns. *Biopolymers* **38:** 705-721.

Peng, T., Zintsmaster, J.S., Namanja, A.T., and Peng, J.W. 2007. Sequence-specific dynamics modulate recognition specificity in WW domains. *Nat Struct Mol Biol* **14:** 325-331.

Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* **277:** 985-994.

Pollastri, G., and Baldi, P. 2002. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* **18 Suppl 1:** S62-70.

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47:** 228-235.

Ponting, C.P., and Russell, R.R. 2002. The natural history of protein domains. *Annu Rev Biophys Biomol Struct* **31:** 45-71.

Pugalenthi, G., Archunan, G., and Sowdhamini, R. 2005. DIAL: a web-based server for the automatic identification of structural domains in proteins. *Nucleic Acids Res* **33:** W130-132.

Punta, M., and Rost, B. 2005. PROFcon: novel prediction of long-range contacts. *Bioinformatics* **21:** 2960-2968.

Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., and Dunker, A.K. 2004. Protein flexibility and intrinsic disorder. *Protein Sci* **13:** 71-80.

Rajashankar, K.R., and Ramakumar, S. 1996. Pi-turns in proteins and peptides: Classification, conformation, occurrence, hydration and sequence. *Protein Sci* **5:** 932-946.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7:** 95-99.

Ramachandran, G.N., and Sasisekharan, V. 1968. Conformation of polypeptides and proteins. *Adv Protein Chem* **23:** 283-438.

Rawn, J.D. 1990. *Traité de Biochimie*, Editions Universitaires ed. Editions Universitaires, Paris.

Receveur-Brechot, V., Bourhis, J.M., Uversky, V.N., Canard, B., and Longhi, S. 2006. Assessing protein disorder and induced folding. *Proteins* **62:** 24-45.

Regad, L., Guyon, F., Maupetit, J., Tufféry, P., and Camproux, A.C. 2008. A hidden Markov Model applied to the protein 3D structure analysis. *Computational Statistics & Data Analysis***:** 3198-3207.

Richards, F.M., and Kundrot, C.E. 1988. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins* **3:** 71-84.

Richardson, J.S., Getzoff, E.D., and Richardson, D.C. 1978. The beta bulge: a common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci U S A* **75:** 2574-2578.

Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q., and Baker, D. 1997. Functional rapidly folding proteins from simplified amino acid sequences. *Nat Struct Biol* **4:** 805-809.

Rogov, S.I., and Nekrasov, A.N. 2001. A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences. *Protein Eng* **14:** 459-463.

Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. 2004. Protein structure prediction using Rosetta. *Methods Enzymol* **383:** 66-93.

Rooman, M.J., Rodriguez, J., and Wodak, S.J. 1990. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol* **213:** 327-336.

Rose, G.D., and Wolfenden, R. 1993. Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu Rev Biophys Biomol Struct* **22:** 381-415.

Rost, B., and Sander, C. 1998. 3rd Generation Prediction Of Secondary Structure. In *Predicting protein structure*. (ed. W.D.M. (ed.)). Humana Press.

Rost, B., Sander, C., and Schneider, R. 1994. PHD--an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci* **10:** 53-60.

Sacquin-Mora, S., Laforet, E., and Lavery, R. 2007. Locating the active sites of enzymes using mechanical properties. *Proteins* **67:** 350-359.

Sacquin-Mora, S., and Lavery, R. 2006. Investigating the local flexibility of functional residues in hemoproteins. *Biophys J* **90:** 2706-2717.

Sali, A., and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234:** 779-815.

Samanta, U., Pal, D., and Chakrabarti, P. 2000. Environment of tryptophan side chains in proteins. *Proteins* **38:** 288-300.

Sander, O., Sommer, I., and Lengauer, T. 2006. Local protein structure prediction using discriminative models. *BMC Bioinformatics* **7:** 14.

Sanejouand, Y.H. 2007. Les modes normaux de vibration de basse frequence des proteines. Universite de Lyon-I.

Sanger, F. 1959. Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes. *Science* **129:** 1340-1344.

Sawada, Y., and Honda, S. 2009. ProSeg: a database of local structures of protein segments. *J Comput Aided Mol Des* **23:** 163-169.

Scapin, G., Gordon, J.I., and Sacchettini, J.C. 1992. Refinement of the structure of recombinant rat intestinal fatty acid-binding apoprotein at 1.2-A resolution. *J Biol Chem* **267:** 4253-4269.

Schlessinger, A., and Rost, B. 2005. Protein flexibility and rigidity predicted from sequence. *Proteins* **61:** 115-126.

Schuchhardt, J., Schneider, G., Reichelt, J., Schomburg, D., and Wrede, P. 1996. Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng* **9:** 833-842.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H.J. 2004. Recognition of functional sites in protein structures. *J Mol Biol* **339:** 607-633.

Sibanda, B.L., and Thornton, J.M. 1985. Beta-hairpin families in globular proteins. *Nature* **316:** 170-174.

Siddiqui, A.S., and Barton, G.J. 1995. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci* **4:** 872-884.

Siddiqui, A.S., Dengler, U., and Barton, G.J. 2001. 3Dee: a database of protein structural domains. *Bioinformatics* **17:** 200-201.

Sklenar, H., Etchebest, C., and Lavery, R. 1989. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins* **6:** 46-60.

Smith, D.K., Radivojac, P., Obradovic, Z., Dunker, A.K., and Zhu, G. 2003. Improved amino acid flexibility parameters. *Protein Sci* **12:** 1060-1072.
Snyder, D.A., and Montelione, G.T. 2005. Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. *Proteins* **59:** 673-686.
Song, J., Burrage, K., Yuan, Z., and Huber, T. 2006. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC Bioinformatics* **7:** 124.
Soto, C.S., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. 2008. Loop modeling: Sampling, filtering, and scoring. *Proteins* **70:** 834-843.
Sowdhamini, R., and Blundell, T.L. 1995. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci* **4:** 506-520.
Srinivasan, R., and Rose, G.D. 1999. A physical basis for protein secondary structure. *Proc Natl Acad Sci U S A* **96:** 14258-14263.
Stryer, L. 1996. *Biochemistry*. Fourth ed. Freeman, W.H. and Company, New York.
Suhre, K., and Sanejouand, Y.H. 2004. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res* **32:** W610-614.
Swindells, M.B. 1995. A procedure for detecting structural domains in proteins. *Protein Sci* **4:** 103-112.
Tartaglia, G.G., Cavalli, A., and Vendruscolo, M. 2007. Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure* **15:** 139-143.
Taylor, W.R. 1986. The classification of amino acid conservation. *J Theor Biol* **119:** 205-218.
Taylor, W.R. 2007. Evolutionary transitions in protein fold space. *Curr Opin Struct Biol* **17:** 354-361.
Thomas, A., Bouffioux, O., Geeurickx, D., and Brasseur, R. 2001. Pex, analytical tools for PDB files. I. GF-Pex: basic file to describe a protein. *Proteins* **43:** 28-36.
Thomas, A., Deshayes, S., Decaffmeyer, M., Van Eyck, M.H., Charloteaux, B., and Brasseur, R. 2006. Prediction of peptide structure: how far are we? *Proteins* **65:** 889-897.
Thomas, A., Meurisse, R., Charloteaux, B., and Brasseur, R. 2002. Aromatic side-chain interactions in proteins. I. Main structural features. *Proteins* **48:** 628-634.
Thornton, J.M., Orengo, C.A., Todd, A.E., and Pearl, F.M. 1999. Protein folds, functions and evolution. *J Mol Biol* **293:** 333-342.
Tironi, I.G., Sperb, R., Smith, P.E., and van Gunsteren, W.F. 1995. Generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* **102:** 5451–5459.
Tompa, P. 2002. Intrinsically unstructured proteins. *Trends Biochem Sci* **27:** 527-533.
Tompa, P. 2008. Prediction of protein disorder. In *Atelier de Formation #185 - Protéines intrinsèquement désordonnées et pathologies associées : prédiction, caractérisation et fonction*, Saint-Raphaël (France).
Travaglini-Allocatelli, C., Ivarsson, Y., Jemth, P., and Gianni, S. 2009. Folding and stability of globular proteins and implications for function. *Curr Opin Struct Biol* **19:** 3-7.
Tsai, C.J., and Nussinov, R. 1997. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci* **6:** 24-42.
Tung, C.H., Huang, J.W., and Yang, J.M. 2007. Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. *Genome Biol* **8:** R31.
Tyagi, M., Bornot, A., Offmann, B., and de Brevern, A.G. 2009a. Analysis of loop boundaries using different local structure assignment methods. *Protein Science, in press*.
Tyagi, M., Bornot, A., Offmann, B., and De Brevern, A.G. 2009b. Protein short loop prediction in terms of a structural alphabet. *Computational Biology and Chemistry* **33:** 329-333.
Tyagi, M., de Brevern, A.G., Srinivasan, N., and Offmann, B. 2008. Protein structure mining using a structural alphabet. *Proteins* **71:** 920-937.
Tyagi, M., Gowri, V.S., Srinivasan, N., de Brevern, A.G., and Offmann, B. 2006a. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins* **65:** 32-39.
Tyagi, M., Sharma, P., Swamy, C.S., Cadet, F., Srinivasan, N., de Brevern, A.G., and Offmann, B. 2006b. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Res* **34:** W119-123.

Unger, R., Harel, D., Wherland, S., and Sussman, J.L. 1989. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* **5:** 355-373.
van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P., and Tironi, I.G. 1996. Biomolecular Simulation: The GROMOS96 Manual and User Guide. 1042.
Venkatachalam, C.M. 1968. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6:** 1425-1436.
Vihinen, M., Torkkila, E., and Riikonen, P. 1994. Accuracy of protein flexibility predictions. *Proteins* **19:** 141-149.
Voet, D., and Voet, J.G. 1995. *Biochemistry*, Second Edition ed. John Wiley & sons, INC, New York, pp. 1361.
Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., et al. 2005. DisProt: a database of protein disorder. *Bioinformatics* **21:** 137-140.
Vullo, A., Walsh, I., and Pollastri, G. 2006. A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics* **7:** 180.
Wang, J., and Wang, W. 1999. A computational approach to simplifying the protein folding alphabet. *Nat Struct Biol* **6:** 1033-1038.
Ward, J.J., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2003. Secondary structure prediction with support vector machines. *Bioinformatics* **19:** 1650-1655.
Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337:** 635-645.
Wei, Y., Kim, S., Fela, D., Baum, J., and Hecht, M.H. 2003. Solution structure of a de novo protein from a designed combinatorial library. *Proc Natl Acad Sci U S A* **100:** 13270-13273.
Weichsel, A., Gasdaska, J.R., Powis, G., and Montfort, W.R. 1996. Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer. *Structure* **4:** 735-751.
Weiss, D.R., and Levitt, M. 2009. Can morphing methods predict intermediate structures? *J Mol Biol* **385:** 665-674.
Wernisch, L., Hunting, M., and Wodak, S.J. 1999. Identification of structural domains in proteins by a graph heuristic. *Proteins* **35:** 338-352.
Wetlaufer, D.B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A* **70:** 697-701.
Wetlaufer, D.B. 1981. Folding of protein fragments. *Adv Protein Chem* **34:** 61-92.
Williamson, M.P., Havel, T.F., and Wuthrich, K. 1985. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *J Mol Biol* **182:** 295-315.
Wintjens, R., Wodak, S.J., and Rooman, M. 1998. Typical interaction patterns in alphabeta and betaalpha turn motifs. *Protein Eng* **11:** 505-522.
Wintjens, R.T., Rooman, M.J., and Wodak, S.J. 1996. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J Mol Biol* **255:** 235-253.
Wodak, S.J., and Janin, J. 1981. Location of structural domains in protein. *Biochemistry* **20:** 6544-6552.
Wojcik, J., Mornon, J.P., and Chomilier, J. 1999. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol* **289:** 1469-1490.
Wouters, M.A., and Curmi, P.M. 1995. An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins* **22:** 119-131.
Wrinch, D. 1940. The fabric structure of proteins with special reference to cytogenetics. *Journal of Genetics* **40:** 359-378.
Xiang, Z., and Honig, B. 2001. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* **311:** 421-430.

Xiang, Z., Soto, C.S., and Honig, B. 2002. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci U S A* **99:** 7432-7437.

Xu, J. 2005. Rapid side-chain prediction via tree decomposition. In *RECOMB*.

Xu, Y., Xu, D., and Gabow, H.N. 2000. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* **16:** 1091-1104.

Yang, A.S., and Wang, L.Y. 2003. Local structure prediction with local structure-based sequence profiles. *Bioinformatics* **19:** 1267-1274.

Yang, J. 2008. Comprehensive description of protein structures using protein folding shape code. *Proteins* **71:** 1497-1518.

Yuan, Z., Bailey, T.L., and Teasdale, R.D. 2005. Prediction of protein B-factor profiles. *Proteins* **58:** 905-912.

Zemla, A., Venclovas, C., Moult, J., and Fidelis, K. 1999. Processing and analysis of CASP3 protein structure predictions. *Proteins* **Suppl 3:** 22-29.

Zhang, H., Zhang, T., Chen, K., Shen, S., Ruan, J., and Kurgan, L. 2009. On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins* **76:** 617-636.

Zhang, L., and Skolnick, J. 1998. How do potentials derived from structural databases relate to "true" potentials? *Protein Sci* **7:** 112-122.

Zhou, H., Xue, B., and Zhou, Y. 2007. DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile. *Protein Sci* **16:** 947-955.

Zhou, H., and Zhou, Y. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* **11:** 2714-2726.

Zhu, K., Pincus, D.L., Zhao, S., and Friesner, R.A. 2006. Long loop prediction using the protein local optimization program. *Proteins* **65:** 438-452.

Zimmermann, O., and Hansmann, U.H. 2008. LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. *J Chem Inf Model* **48:** 1903-1908.